OM of: US-09-303-518D-465 to: SwissProt_40:* out_format : pfs

Date: Jun 30, 2002 8:40 AM

About: Results were produced by the GenCore software, version 4.5,
Copyright (c) 1993-2000 Compugen Ltd.

Command line parameters:
-MODEL=frame+_n2p.model -DEV=xlh
-Q=/cgn2_1/USPTO_spool/US09303518/runat_28062002_142714_4362/app_query.fasta_1.23501
-DB=SwissProt_40 -QFMT=fastan -SUFFIX=rsp -GAPOP=12.000
-GAPEXT=4.000 -MINMATCH=0.100 -LOOPCL=0.000 -LOOPEXT=0.000
-QGAPOP=4.500 -QGAPEXT=0.050 -XGAPOP=10.000 -XGAPEXT=0.500
-FGAPOP=6.000 -FGAPEXT=7.000 -YGAPOP=10.000 -YGAPEXT=0.500
-DELOP=6.000 -DELEXT=7.000 -START=1 -MATRIX=blosum62
-TRANS=human40.cdi -LIST=100 -DOCALIGN=200 -THR_SCORE=pct
-THR_MAX=100 -THR_MIN=0 -ALIGN=45 -MODE=LOCAL -OUTFMT=pfs
-NORM=ext -HEAPSIZE=500 -MINLEN=0 -MAXLEN=2000000000
-USER=US09303518_@CGN1_1_440 -NCPU=6 -ICPU=3 -LONGLOG
-DEV_TIMEOUT=120 -WARN_TIMEOUT=30 -NO_XLPXY -WAIT -THREADS=1

Search information block:
Query: US-09-303-518D-465
Query length: 1671
Database: SwissProt_40:*
Database sequences: 105224
Database length: 38719550
Search time (sec): 217.960000

score_list:

Sequence	Strd	Orig	ZScore	EScore	Len	Documentation
SwissProt_40:MUC2_HUMAN	+	153.50	148.97	0.0258	5179	Q02817 homo sapiens (human)
SwissProt_40:YWR1_CAEEL	+	148.50	156.15	0.0434	1223	Q10925 caenorhabditis elega
SwissProt_40:AMYH_YEAST	+	147.00	153.46	0.0549	1367	P08640 saccharomyces cerevi
SwissProt_40:DAN4_YEAST	+	141.00	148.12	0.1281	1161	P47179 saccharomyces cerevi
SwissProt_40:GTFC_STRMU	+	140.00	145.49	0.1517	1375	P13470 streptococcus mutans
SwissProt_40:DRPL_RAT	+	139.00	145.69	0.1717	1183	P54258 rattus norvegicus (rat
SwissProt_40:VRP1_YEAST	+	136.50	146.16	0.2343	817	P37370 saccharomyces cerevis
SwissProt_40:YGY3_HALSQ	+	135.00	150.03	0.2667	437	P21561 haloferax sp. (strain
SwissProt_40:MUC1_MESAU	+	132.50	143.32	0.4076	676	Q60528 mesocricetus auratus
SwissProt_40:VGP3_EBVA8	+	131.50	139.78	0.4896	886	Q07284 epstein-barr virus (s
SwissProt_40:YAV1_SCHPO	+	130.50	132.37	0.6253	1794	Q10172 schizosaccharomyces
SwissProt_40:DRPL_HUMAN	+	130.00	135.49	0.6341	1185	P54259 homo sapiens (human)
SwissProt_40:VGLX_HSVEB	+	128.50	137.32	0.7455	797	P28968 equine herpesvirus ty
SwissProt_40:FIG2_YEAST	+	128.00	130.51	0.8851	1609	P25653 saccharomyces cerevi
SwissProt_40:ENAM_PIG	+	127.00	132.43	0.9751	1142	O97939 sus scrofa (pig). enam
SwissProt_40:KN8R_YEAST	+	126.50	134.05	1.01	893	P53739 saccharomyces cerevis
SwissProt_40:GTFB_STRMU	+	125.50	128.45	1.26	1476	P08987 streptococcus mutans
SwissProt_40:SON_HUMAN	+	125.50	124.03	1.35	2426	P18583 homo sapiens (human).
SwissProt_40:PHP_DROME	+	125.00	127.23	1.37	1589	P39769 drosophila melanogast
SwissProt_40:A180_MOUSE	+	124.00	131.14	1.46	901	Q61548 mus musculus (mouse).
SwissProt_40:YM96_YEAST	+	123.50	128.48	1.62	1140	Q04893 saccharomyces cerevi
SwissProt_40:VGP3_EBV	+	123.00	129.95	1.69	907	P03200 epstein-barr virus (str
SwissProt_40:PODX_RABIT	+	121.50	132.69	1.95	551	Q28645 oryctolagus cuniculus
SwissProt_40:EGR1_BRARE	+	121.00	132.79	2.08	511	P26632 brachydanio rerio (ze
SwissProt_40:HAP1_HAEIN	+	121.00	123.77	2.40	1409	P44596 haemophilus influenz
SwissProt_40:YB23_HUMAN	+	120.50	128.60	2.37	768	Q9ulj7 homo sapiens (human).
SwissProt_40:P115_CHICK	+	120.00	128.10	2.54	762	Q98917 gallus gallus (chicke
SwissProt_40:SON_MOUSE	+	120.00	117.88	2.99	2404	Q9qx47 mus musculus (mouse).
SwissProt_40:YS8A_CAEEL	+	119.00	126.58	2.96	796	Q09625 caenorhabditis elegan
SwissProt_40:A180_RAT	+	119.00	125.35	3.02	915	Q05140 rattus norvegicus (rat)
SwissProt_40:NRG3_MOUSE	+	118.50	127.00	3.13	713	O35181 mus musculus (mouse).
SwissProt_40:FYB_MOUSE	+	118.50	125.77	3.20	819	O35601 mus musculus (mouse).
SwissProt_40:CAPP_MYCLE	+	118.00	124.03	3.50	934	P46710 mycobacterium leprae.
SwissProt_40:SYJ1_HUMAN	+	118.00	119.38	3.77	1575	O43426 homo sapiens (human)
SwissProt_40:TFE2_HUMAN	+	117.50	126.64	3.58	654	P15923 homo sapiens (human).
SwissProt_40:KIN4_YEAST	+	117.50	124.84	3.68	800	Q01919 saccharomyces cerevis
SwissProt_40:CHI2_COCIM	+	117.50	124.20	3.72	860	P54197 coccidioides immitis.
SwissProt_40:ITF2_MOUSE	+	117.00	125.85	3.86	670	Q60722 mus musculus (mouse).
SwissProt_40:TEGU_HSVEB	+	117.00	111.35	4.86	3421	P28955 equine herpesvirus t
SwissProt_40:MUC1_XENLA	+	116.50	125.40	4.15	662	Q05049 xenopus laevis (afrid
SwissProt_40:SUZ2_DROME	+	115.50	117.83	5.31	1365	P25172 drosophila melano
SwissProt_40:OMPA_RICCN	+	115.00	113.77	6.03	2021	Q52657 rickettsia conori
SwissProt_40:LG1_MAIZE	+	114.50	127.64	5.16	399	O04003 zea mays (maize). l
SwissProt_40:MAP4_MOUSE	+	114.50	118.41	5.97	1125	P27546 mus musculus (mou
SwissProt_40:NCR2_HUMAN	+	114.50	111.25	6.69	2517	Q9y618 h nuclear recepto
SwissProt_40:NRG3_HUMAN	+	114.00	121.82	6.03	720	P56975 homo sapiens (huma
SwissProt_40:VE2_HPV25	+	113.50	124.46	6.16	502	P36787 human papillomaviru
SwissProt_40:PLD1_CRIGR	+	113.50	118.02	6.82	1036	O08684 cricetulus griseu
SwissProt_40:NCR2_MOUSE	+	113.50	110.28	7.71	2472	Q9wu42 mus musculus nucl
SwissProt_40:HSP2_MACNE	+	113.00	137.99	5.30	103	P35298 macaca nemestrina
SwissProt_40:WETA_PENCH	+	113.00	123.95	6.62	499	Q01870 penicillium chryso
SwissProt_40:STUA_EMENI	+	113.00	122.46	6.78	590	P36011 emericella nidulan
SwissProt_40:VE2_HPV08	+	112.50	123.40	7.12	498	P06422 human papillomaviru
SwissProt_40:MUC4_HUMAN	+	112.50	121.60	7.32	610	Q99102 homo sapiens (huma
SwissProt_40:MAP4_HUMAN	+	112.50	115.94	8.01	1152	P27816 homo sapiens (hum
SwissProt_40:CHI3_CANAL	+	112.00	121.68	7.79	567	P40954 candida albicans (
SwissProt_40:YQU3_CAEEL	+	112.00	114.64	8.71	1251	Q09550 caenorhabditis el
SwissProt_40:BAI2_HUMAN	+	112.00	112.61	9.00	1572	O60241 homo sapiens (hum
SwissProt_40:PSC_DROME	+	112.00	112.44	9.02	1603	P35820 drosophila melanog
SwissProt_40:RPB1_DROME	+	112.00	110.98	9.23	1887	P04052 drosophila melano
SwissProt_40:CCAA_MOUSE	+	112.00	109.77	9.41	2164	P97445 mus musculus (mou
SwissProt_40:SFRA_MOUSE	+	111.50	127.30	7.60	283	Q60701 mus musculus (mous
SwissProt_40:EGR1_HUMAN	+	111.50	121.50	8.33	543	P18146 homo sapiens (huma
SwissProt_40:HWP1_CANAL	+	111.50	120.12	8.51	634	P46593 candida albicans (
SwissProt_40:YNR6_YEAST	+	111.50	120.09	8.52	636	P53882 saccharomyces cere
SwissProt_40:YAG3_YEAST	+	111.50	113.58	9.44	1322	P39712 saccharomyces cer
SwissProt_40:IE18_PRVKA	+	111.50	112.79	9.56	1446	P33479 pseudorabies viru
SwissProt_40:CRK7_HUMAN	+	111.50	112.52	9.60	1490	Q9nyv4 homo sapiens (hum
SwissProt_40:FLO1_YEAST	+	111.50	112.24	9.64	1537	P32768 saccharomyces cer
SwissProt_40:TFE2_RAT	+	111.00	119.35	9.18	649	P21677 rattus norvegicus (r
SwissProt_40:YK82_MYCTU	+	110.50	117.85	10.02	721	Q10690 mycobacterium tube
SwissProt_40:PGCA_RAT	+	110.50	108.23	11.67	2124	P07897 rattus norvegicus (
SwissProt_40:CCAA_RAT	+	110.50	107.87	11.74	2212	P54282 rattus norvegicus (
SwissProt_40:CREA_ASPNG	+	110.00	121.94	10.01	427	Q05620 aspergillus niger.
SwissProt_40:YHC8_YEAST	+	110.00	118.84	10.51	605	P38739 saccharomyces cere
SwissProt_40:MRP_STRSU	+	110.00	112.34	11.65	1256	P32653 streptococcus suis
SwissProt_40:PMPG_CHLTR	+	109.50	113.69	12.16	1013	O84879 chlamydia trachom
SwissProt_40:MANA_RHOMR	+	109.50	113.62	12.17	1021	P49425 rhodothermus mari
SwissProt_40:ZO2_MOUSE	+	109.50	112.43	12.40	1167	Q9z0u1 mus musculus (mous
SwissProt_40:IRS1_HUMAN	+	109.50	111.88	12.51	1242	P35568 homo sapiens (hum
SwissProt_40:FSH_DROME	+	109.50	107.47	13.42	2038	P13709 drosophila melanog
SwissProt_40:P73_HUMAN	+	109.00	117.26	12.24	636	O15350 homo sapiens (human
SwissProt_40:DAB_DROME	+	109.00	105.41	14.77	2411	P98081 drosophila melanog
SwissProt_40:LEUK_HUMAN	+	108.50	120.82	12.33	400	P16150 homo sapiens (huma
SwissProt_40:CAUP_DROME	+	108.50	115.93	13.32	693	P54269 drosophila melanog
SwissProt_40:KP58_DROME	+	108.50	113.11	13.93	952	Q9vpc0 drosophila melanog
SwissProt_40:N121_HUMAN	+	108.50	110.84	14.44	1229	Q9y2n3 homo sapiens (hum
SwissProt_40:SOL_DROME	+	108.50	108.51	14.99	1597	P27398 drosophila melanog
SwissProt_40:PKSK_BACSU	+	108.50	99.40	17.32	4447	P40803 bacillus subtilis
SwissProt_40:ATS4_HUMAN	+	108.00	113.69	14.71	837	O75173 homo sapiens (huma
SwissProt_40:LUMI_ARATH	+	108.00	112.53	14.99	953	Q38796 arabidopsis thalia
SwissProt_40:ITA4_HUMAN	-	108.00	111.77	15.17	1038	P13612 homo sapiens (hum
SwissProt_40:IGA_NEIGO	+	108.00	108.31	16.02	1532	P09790 neisseria gonorrho
SwissProt_40:VE2_HPV5B	+	107.50	117.46	14.77	514	P26545 human papillomaviru
SwissProt_40:GUXA_CELFI	-	107.50	112.76	15.91	872	P50401 cellulomonas fimi.
SwissProt_40:UTX_MOUSE	+	107.50	108.98	16.89	1333	O70546 mus musculus (mous
SwissProt_40:TSC2_RAT	+	107.50	106.27	17.64	1809	P49816 rattus norvegicus (
SwissProt_40:N214_HUMAN	+	107.50	104.98	18.00	2090	P35658 homo sapiens (hum
SwissProt_40:CCAA_HUMAN	+	107.50	103.37	18.46	2505	O00555 homo sapiens (hum
SwissProt_40:HSP2_MOUSE	+	107.00	130.86	12.73	107	P07978 mus musculus (mous

seq_name: SwissProt_40:MUC2_HUMAN

seq_documentation_block:
ID MUC2_HUMAN STANDARD; PRT; 5179 AA.
AC Q02817; Q14878;
DT 01-JUN-1994 (Rel. 29, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 16-OCT-2001 (Rel. 40, Last annotation update)
DE Mucin 2 precursor (Intestinal mucin 2).
GN MUC2 OR SMUC.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

```
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
```

FT	SIGNAL	1	20	POTENTIAL.
FT	CHAIN	21	5179	MUCIN 2.
FT	DOMAIN	1401	1747	APPROXIMATE REPEATS.
FT	REPEAT	1401	1416	1.
FT	REPEAT	1417	1432	2.
FT	REPEAT	1433	1448	3.
FT	REPEAT	1449	1464	4.
FT	REPEAT	1465	1471	5.
FT	REPEAT	1472	1478	6.
FT	REPEAT	1479	1494	7A.
FT	REPEAT	1495	1517	7B.
FT	REPEAT	1518	1533	8A.
FT	REPEAT	1534	1556	8B.
FT	REPEAT	1557	1572	9A.
FT	REPEAT	1573	1596	9B.
FT	REPEAT	1597	1612	10A.
FT	REPEAT	1613	1635	10B.
FT	REPEAT	1636	1651	11A.
FT	REPEAT	1652	1675	11B.
FT	REPEAT	1676	1683	12.
FT	REPEAT	1684	1699	13.
FT	REPEAT	1700	1715	14.
FT	REPEAT	1716	1731	15.
FT	REPEAT	1732	1747	16.
FT	DOMAIN	4815	4886	VWFC 1.
FT	DOMAIN	4924	4991	VWFC 2.
FT	DOMAIN	5075	5160	CTCK.
FT	DISULFID	5075	5122	BY SIMILARITY.
FT	DISULFID	5089	5136	BY SIMILARITY.
FT	DISULFID	5098	5152	BY SIMILARITY.
FT	DISULFID	5102	5154	BY SIMILARITY.
FT	DISULFID	?	5159	BY SIMILARITY.
FT	CARBOHYD	163	163	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	423	423	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	670	670	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	770	770	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	894	894	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1139	1139	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1154	1154	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1215	1215	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1230	1230	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1246	1246	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1787	1787	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	1820	1820	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4339	4339	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4351	4351	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4362	4362	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4373	4373	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4422	4422	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4438	4438	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4502	4502	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4616	4616	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4627	4627	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4752	4752	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4787	4787	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4881	4881	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4888	4888	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4955	4955	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	4970	4970	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	5019	5019	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	5038	5038	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CARBOHYD	5069	5069	N-LINKED (GLCNAC. . .) (POTENTIAL).
FT	CONFLICT	1351	1351	H -> L (IN REF. 3).
FT	CONFLICT	1412	1412	T -> S (IN REF. 3).
FT	CONFLICT	1449	1449	L -> P (IN REF. 3).
FT	CONFLICT	1504	1504	M -> T (IN REF. 3).
FT	CONFLICT	4192	4192	G -> S (IN REF. 2).
SQ	SEQUENCE	5179 AA;	540295 MW;	85CD7571FB9A5663 CRC64;

```
alignment_scores:
          Quality:  153.50          Length:    315
            Ratio:  1.104            Gaps:     13
```

```
 Percent Similarity:  44.127   Percent Identity: 24.444

alignment_block:
 US-09-303-518D-465 x MUC2_HUMAN   ..

 Align seg 1/1  to: MUC2_HUMAN  from: 1  to: 5179

     269 ACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGAC 318
         |||   :::      |||   |||   |||   :::   ||||||   ||
    1520 ThrThrThrProSerProProThrThrThrThrThrProProProTh 1536

     319 AACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGG 368
         ||||  |||      |||           |||:::::|||   |||:
    1536 rThrThrProSer...ProProThrThrThrProIleThrProProThrS 1552

     369 ATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCG 418
         :::::      |||:::         |||   |||   :::   ||||||
    1552 erThr......ThrThrLeuProProThrThrThrProSerProProPro 1566

     419 ACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCG 468
         |||:::  |||            :::::  |||  |||
    1567 ThrThrThrThrThrProProProThrThrThrProSerPro........ 1580

     469 AGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCT 518
                 :::||||||     :::  |||  |||:::::::
    1581 ..........ProThrThrThrThrProSerProProThrIleThrThrT 1594

     519 CAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCC 568
         ::|||  |||  |||:::::::|||:::        :::|||::::::
    1594 hrThrProProProThrThrThrProSerProProThrThrThrThrThr 1610

     569 ACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGC 618
         |||  |||                          |||  :::  ::
    1611 ThrProPro.......................ProThrThrThrProSe 1619

     619 GCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGC 668
         :||||||  |||:::|||                  ||||||
    1619 rProProThrThrThrPro..............IleThrProProThrS 1631

     669 TTTCAACGGCACTGCAGATATCGTCA..............AAAACATCA 703
           :::|||:::|||      ::::::                |||:::
    1631 erThrThrThrLeuProProThrThrThrProSerProProProThrThr 1647

     704 TCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATA 753
         ::::::               ::::::  :::  |||      :::
    1648 ThrThrThrProProProThrThrThrProSerProProThrThrThrTh 1664

     754 AGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCAC 803
           :::                  |||                 ||||
    1664 rProSerProProIleThrThrThrThrThr..............ProP 1676

     804 CGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTCA 853
         ||  |||        :::|||  |||
    1676 roProThrThrThrProSerSerProIle..................... 1685

     854 AAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAAT 903
           |||  |||:::  |||::::::              ||||||
    1686 ThrThrThrProSerProProThrThr......ThrMetThrThrProSe 1700

     904 GCCGCACAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCC 953
          |||      :::     ||||||:::                :::::|
    1700 rProThrThrThr......ProSerSerProIleThrThrThrThrThrP 1715

     954 CGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCA 1003
         ||||||:::                  |||:::  ::::::
    1715 roSerSerThrThrThrProSerProProProThrThrMetThrThrPro 1731

    1004 CGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAA 1053
                                        |||
    1732 ................................SerProThrThrTh 1736

    1054 GGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCAAATA 1103
           :::||||||::::::|||:::      :::      |||  ::::

    1736 rProSerProProThrThrThrMetThrThrLeuProProThrThrThrS 1753

    1104 CCCGTCCCCTTACCATT.....CCCGAAATATCCGTTCAAACT 1141
         ::  ||||||||          |||:::|||:::     |||
    1753 erSerProLeuThrThrThrProLeuProProSerIleThrProProThr 1769

    1142 TGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGC 1186
         :::|||      ::::::      |||:::|||      ||||||
    1770 PheSerProPheSerThr......ThrThrProThrThrProCys 1782

seq_name: SwissProt_40:YWR1_CAEEL

seq_documentation_block:
ID   YWR1_CAEEL     STANDARD;      PRT;  1223 AA.
AC   Q10925;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Putative tyrosine-protein kinase B0302.1 in chromosome X
DE   (EC 2.7.1.112).
GN   B0302.1.
OS   Caenorhabditis elegans.
OC   Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Rhabditoidea;
OC   Rhabditidae; Peloderinae; Caenorhabditis.
OX   NCBI_TaxID=6239;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=BRISTOL N2;
RA   Du Z.;
RL   Submitted (NOV-1995) to the EMBL/GenBank/DDBJ databases.
CC   -!- CATALYTIC ACTIVITY: ATP + a protein tyrosine = ADP + protein
CC       tyrosine phosphate.
CC   -!- SIMILARITY: TO OTHER PROTEIN-TYROSINE KINASES IN THE CATALYTIC
CC       DOMAIN. HIGHEST TO THE SYK/ZAP-70 SUBFAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U41032; AAA82367.1; -.
DR   HSSP; P11362; 1AGW.
DR   WormPep; B0302.1; CE03866.
DR   InterPro; IPR000719; Euk_pkinase.
DR   InterPro; IPR001452; SH3.
DR   InterPro; IPR001245; Tyr_pkinase.
DR   Pfam; PF00069; pkinase; 1.
DR   Pfam; PF00018; SH3; 1.
DR   PRINTS; PR00109; TYRKINASE.
DR   SMART; SM00326; SH3; 1.
DR   SMART; SM00219; TyrKc; 1.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; 1.
DR   PROSITE; PS00109; PROTEIN_KINASE_TYR; 1.
DR   PROSITE; PS50011; PROTEIN_KINASE_DOM; 1.
KW   Hypothetical protein; Transferase; Tyrosine-protein kinase;
KW   ATP-binding.
FT   DOMAIN      107    369       PROTEIN KINASE.
FT   NP_BIND     113    121       ATP (BY SIMILARITY).
FT   BINDING     139    139       ATP (BY SIMILARITY).
FT   ACT_SITE    230    230       BY SIMILARITY.
SQ   SEQUENCE   1223 AA;  132940 MW;  19EFFD8E0A815227 CRC64;


alignment_scores:
            Quality:  148.50             Length:   530
              Ratio:  0.640                Gaps:    23
 Percent Similarity:  43.774   Percent Identity: 22.453

alignment_block:
 US-09-303-518D-465 x YWR1_CAEEL   ..
```

```
Align seg 1/1  to: YWR1_CAEEL  from: 1  to: 1223

 12 CCGCAAAATATCCCTTATTCTGTCCATACTGGCAGTGTGCCTGCCGATGC 61
    |||   |||||||||      |||  |||   :::   ||||||   ||
715 ProAlaAsnIleProCysLeuValProThrProAlaProProAlaProAl 731

 62 ATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTC 111
    |                        :::|||      |||::::::|||:
731 a.....................HisPheSerGlnProValSerSerG 740

112 GACCGTCAGCATT................................... 124
    ::      |||
740 lnArgValAlaGlnGlnGlnGlnAsnThrLeuGlnLysAlaLeuAsnAsp 756

125 ....................TCGAACCCGACGGGAAATACCACCTATTC 153
                        ::::::|||||||||:::|||::::
757 GluLeuLysGlyAsnLeuAsnLysArgProThrGlyThrThrAlaProPr 773

154 GGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAA 203
     :::   |||   |||   |||   |||   |||           :::
773 oSerAsnGlyPheAsnAlaProArgAlaAspValAlaProValGlnGln. 789

204 CATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGG...CGGCCA 250
        :::   ||||||   |||:::      :::   :::   :::|||
790 .....ArgProIleSerSerAlaSerIleProAlaLeuGlnProGlnPro 804

251 TTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAA 300
    :::::::::|||:::      ::::::      :::
805 IleGlnHisIleGlnLysProIleGlnProGlnGlnVal........... 817

301 GTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGC 350
        |||||||||||||||:::      :::   |||:::   ::::::
818 .ArgIleProProSerThrAlaProValGlnLysProValGlnValSerA 834

351 CG..............GTAGTCCCGTTGACGGATTCAGCCTTTACCGCA 385
                    |||:::|||   |||:::|||      :::|||
834 laProThrHisSerAsnValAlaProThrThrSerSerGlnAlaSerAla 850

386 TCCATTGGGACGGATACGAACACCATCCCGCCG.................. 418
    :::               |||   :::||||||
851 AspAla.........ArgAsnProLeuProProLysThrSerProProVa 864

419 ...............ACGGCTATGACGGGCCACAGGGCGGCGGCTATCC 452
                    |||:::         |||   |||   :::
864 lSerAsnThrProIleThrValAlaProValHisAlaAlaProThrThrS 881

453 CGCTCCCAAAGGCGCGAGGGATATAT.........ACAGCTACGACATAA 493
        |||::::::      :::           |||:::||||||
881 erAlaProSerThrSerValValThrArgArgProThrSerThrThrAla 897

494 AAG.............................................. 496
    :::
898 GlnMetSerAspGluGluArgArgSerArgIleAlaMetAspIleSerSe 914

497 .GCGTTGCCCAAAATATCCGCC.........TCAACCTGACCGACAACCG 536
     |||||||||      ||||||         ::::::      |||::::
914 rAlaLeuProAlaProSerAlaLeuLeuTyrGlySerAsnSerThrSerS 931

537 CAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGC 586
    ::   |||:::   ::::::||||||:::|||::::::|||
931 erLeuProSerAlaAlaValSerThrAlaSerSerValPro......... 944

587 TGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGCCCC 636
                  :::|||      :::   |||   :::|||   ||
945 ..............SerThrAlaArgAspAsnProValGluThrArgPr 956

637 GAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGA 686
    ||||           :::||||||||||||   |||:::      :
956 oSerGlnProHisValThrMetProProLysLysSerSerGluProIleL 973

687 TATCGTCA.................AAAACATCATCGGCGCGGCAGGAG 718
    ::|||||                 :::      ||||||      :::

973 euSerSerGluValLeuGlnProThrArgLeuProSerAlaThrThrSer 989

719 AAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAAC 768
    :::   :::      ::::::|||   |||      :::         ::
990 GlnAlaLysProValThrGlnProIleArgHisProSerProProValAl 1006

769 ATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGC 818
    :   ::::::   ||||||   |||          |||   ::::::   :
1006 aThrValIleProThrAlaValValAspLysLysProValSerGlnAsnG 1023

819 GCGCATCAACGATT......TGGCAGATATGGCGCAACTCAAAGACTATG 862
    :::::|||::::::      ::::::|||   ::::::|||:::
1023 lnGlySerAsnValProLeuPheAsnIleThrAsnSerSerAsnGlyTyr 1039

863 CCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAA 912
    ||||||                                :::      |||:::
1040 ProGln.....................LeuAsnGlyTyrProAsnTy 1048

913 GGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGG 962
     ::: ::::::   |||   |||   ||||||   :::   :::||| ||
1048 rGly.AsnGlyPheGlnAlaTyrGlyTyrGlyMetAsnTyrHisGln.Gl 1064

963 GATTGGAGCTGTTCGGGGA......AAATACGGCTTGGGCGGCATCACGG 1006
    |       :::   :::|||      :::||||||   |||
1064 yTyrProGlyTyrGlnGlyTyrAsnSerTyrGlyAsnGly.......... 1077

1007 CACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGG 1056
                                |||||:::::||||||      :::
1078 .....................MetGlyGlnLeuAlaLeuThrHisAsn 1086

1057 AAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCAAATACCC 1106
                                        |||   ::::::   ||
1087 ..............................AlaValThrSerLeuPr 1092

1107 GTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACG 1156
     |            |||:::|||      |||:::      |||      |
1092 oProLeuValProSerGluAsnArgPheSerGlyThrAlaGlnProLeuG 1109

1157 GCAAAGAAAACATC.................................. 1170
     ||::::::::::|||
1109 lyGluSerAspIleMetGluPheLeuGlyThrGlnGlnArgGlnAlaGly 1125

1171 ...ACCTCCTCAACCGTGCCGCCGTCAAACGGAAAG.............. 1203
        :::|||   :::||||||||||:::::::::::::
1126 SerSerSerArgAlaValProProAlaSerAlaSerThrSerAlaAlaSe 1142

1204 ..........AATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAG 1243
                :::::::::::::|||:::|||   :::        :::
1142 rGlyIleThrAspLeuSerMetAlaAspLysMetGluValLeuTyrArg. 1158

1244 TGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATAC 1293
                                        |||   |||      :::
1159 ..............................GluAlaAspPheThrHis 1164

1294 GATACGAGAATTAATACCGCTGTACCACAAGTGAAT 1329
           :::   :::|||   |||   |||   |||
1165 LysGlyAsnCysAspThrMetValSerGlnCysAsn 1176
```

```
seq_name: SwissProt_40:AMYH_YEAST

seq_documentation_block:
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Glucoamylase S1/S2 precursor (EC 3.2.1.3) (Glucan 1,4-alpha-
DE   glucosidase) (1,4-alpha-D-glucan glucohydrolase).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
```

```
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: Hydrolysis of terminal 1,4-linked alpha-D-
CC       glucose residues successively from non-reducing ends of the chains
CC       with release of beta-D-glucose.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;


alignment_scores:
            Quality:  147.00              Length:    532
              Ratio:  0.546                 Gaps:     18
 Percent Similarity:  50.564   Percent Identity: 20.865

alignment_block:
 US-09-303-518D-465 x AMYH_YEAST   ..

 Align seg 1/1  to: AMYH_YEAST  from: 1  to: 1367

     110 TCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGC 159
         :::|||   |||   |||   |||:::   :::::|||   ::::::::
     438 ThrThrGluSerSerSerAlaProValThrSerSerThrThrGluSerSe 454

     160 AGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACA 209
         :::::::   :::|||:::   :::      :::      ::::::

     454 rSerAlaProValProThrProSerSerSerThrThrGluSerSerSerA 471

     210 AAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAA 259
             ::::::   :::|||   :::||||||      |||:::
     471 laProValThrSerSerThrThrGluSerSerAlaProValProThr 487

     260 ATATCGGCTACATTG.........TCCGCTTTTCCGATCACGGGCACGAA 300
         |||::::::              |||:::   |||:::|||::::::
     488 ProSerSerSerThrThrGluSerSerSerAlaProValThrSerSerTh 504

     301 GTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGC 350
          :::         |||:::   :::|||                     |
     504 rThrGluSerSerSerAlaProValPro.................ThrP 515

     351 CGGTAGTCCCGTTGACGGATTCAGCCTTTACC..............GCA 385
         ||          |||:::|||:::   :::                 :::
     515 roSerSerSerThrThrGluSerSerSerAlaProAlaProThrProSer 531

     386 TCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCA 435
         |||      |||:::::::::::::   |||   |||:::   |||   ::
     532 SerSerThrThrGluSerSerSerAlaProValThrSerSerThrThrGl 548

     436 CAGGGCGGCGGCT...ATCCCGCTCCCAAAGGCGCGAGGGATATATACAG 482
         :   ::::::|||   :::|||   |||:::                ::::
     548 uSerSerSerAlaProValProThrProSer..............SerS 560

     483 CTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACA 532
         ::||||||   ::::::         :::::::::||||||      :::
     560 erThrThrGluSerSerSerThrProValThrSerSerThrThrGluSer 576

     533 ACCGCAGCA............CCGGACAACGGCTTGTCGACCGTTTCCAC 570
         ::::::|||           |||:::::::::   :::      |||::
     577 SerSerAlaProValProThrProSerSerSerThrThrGluSerSerSe 593

     571 AATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGC 620
         :   ||||||            ::::::   :::   :::|||:::||||
     593 rAlaProValProThrProSerSerSerThrThrGluSerSerSerAlaP 610

     621 CA...CCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAG 667
         ||   |||      :::   |||   |||:::   :::      |||
     610 roAlaProThrProSerSerSerThrThrGluSerSerSerAlaProVal 626

     668 CTTTCAACGGCACTGCAGATATCGTCA.........AAAACATCATCGGC 708
             |||::::::   :::   ||||||            |||   |||::
     627 ThrSerSerThrThrGluSerSerSerAlaProValProThrProSerSe 643

     709 GCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGA 758
         :         :::   |||:::      :::|||               :
     643 rSerThrThrGluSerSerSerAlaProValProThr.........ProS 657

     759 AGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAA 808
         ::::::::|||           :::|||   |||      |||   :::
     657 erSerSerThrThrGluSerSerSerAlaProValProThrProSerSer 673

     809 ACAAGATGGCGCGCATCAACG......ATTTGGCAGATATGGCGCAACTC 852
         :::      ::::::|||:::      :::   :::         :::||
     674 SerThrThrGluSerSerSerAlaProValThrSerSerThrThrGluSe 690

     853 AAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAA 902
         |::::::   |||         |||:::      :::|||::::::|||:
     690 rSerSerAlaProValThrSerSerThrThrGluSerSerSerAlaProV 707

     903 TGCCG.......................CACAAGGCATAGAAGCCG 925
         ::|||                          :::|||      |||
     707 alProThrProSerSerSerThrThrGluSerSerSerAlaProValPro 723

     926 TCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTT 975
         :::      |||         :::|||   |||      :::
     724 ThrProSerSerSerThrThrGluSerSerSerAlaProValProThrPr 740

     976 CGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTC 1025
          :::::::::|||:::   :::::|||      :::   ||||||
     740 oSerSerSerThrThrGluSerSerSerAlaProValThrSerSer.... 755
```

```
 1026 GCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACA 1075
                                    :::      |||::::::
  756 ............................ThrThrGluSerSerSerAla 762

 1076 ATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCCCGA 1125
        :::|||         ||||||::::::                  ::
  763 ProValPro.........ThrProSerSerSerThrThrGluSerSerSe 776

 1126 AATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTC 1175
       :      |||   |||   ||||||   |||:::::::::::::|||   |
  776 rAlaProValProThrProSerSerSerThrThrGluSerSerSerAlaP 793

 1176 CTCAACCGTG.CCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAA 1224
       ||   |||   |||   |||:::      :::   ::::::|||
  793 roValProThrProSerSerSerThrThrGluSerSerValAlaProVal 809

 1225 CGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTT 1274
             |||::::::
  810 ProThrProSerSer..................................... 814

 1275 TGAAAAAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAG 1324
                      ::::::::::|||:::::::::::::|||:::
  815 ..................SerSerAsnIleThrSerSerAlaProSerS 825

 1325 TGAATCCTATAGATGAA.....................CCCGTCTTTAAT 1353
        :::|||::::::::::                     ||||||   :::
  825 erThrProPheSerSerSerThrGluSerSerSerValProValProThr 841

 1354 CCTAAAGGTTCTGTC.........GGATCGGCTCATTCTTGGTCTATAAC 1394
       |||:::::::|||:::         :::||||||          |||   ||
  842 ProSerSerSerThrThrGluSerSerSerAlaProValSerSerSerTh 858

 1395 TGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATA 1444
       |::::::   :::   |||   :::|||      :::
  858 rThrGluSerSerValAlaProValProThrProSer.............. 870

 1445 TCCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAAT 1494
              :::|||   :::   |||||||||      :::   |||
  871 ..SerSerSerAsnIleThrSerSerAlaPro...SerSerIleProPhe 885

 1495 AATGGATATTTGGATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAG 1544
       ::::::      :::::::|||::::::      |||      ||||||
  886 SerSerThrThrGluSerPheSerThrGlyThrThrValThrProSerSe 902

 1545 AACTAAAGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACA 1590
       :::|||            :::   ::::::::::::|||:::|||
  902 rSerLysTyrProGlySerGlnThrGluThrSerValSerSerThr 917
```

seq_name: SwissProt_40:DAN4_YEAST

seq_documentation_block:

```
ID   DAN4_YEAST     STANDARD;      PRT;  1161 AA.
AC   P47179;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Cell wall protein DAN4 precursor.
GN   DAN4 OR YJR151C OR J2223.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RA   Scarcez T.;
RL   Submitted (SEP-1995) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   REGULATION.
RX   MEDLINE=21113168; PubMed=11160904;
RA   Cohen B.D., Sertil O., Abramova N.E., Davies K.J., Lowry C.V.;
RT   "Induction and repression of DAN1 and the family of anaerobic
RT   mannoprotein genes in Saccharomyces cerevisiae occurs through a
RT   complex array of regulatory sites.";
RL   Nucleic Acids Res. 29:799-808(2001).
CC   -!- FUNCTION: COMPONENT OF THE CELL WALL (By similarity).
CC   -!- SUBCELLULAR LOCATION: Attached to the membrane by a GPI-anchor
CC       (Potential).
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED (POTENTIAL).
CC   -!- SIMILARITY: BELONGS TO THE SRP1 / TIP1 FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z49651; CAA89684.1; -.
DR   SGD; S0003912; DAN4.
DR   InterPro; IPR000992; SRP1_TIP1.
DR   Pfam; PF00660; SRP1_TIP1; 1.
DR   PROSITE; PS00724; SRP1_TIP1; 1.
KW   Cell wall; Glycoprotein; Membrane; GPI-anchor; Signal.
FT   SIGNAL        1     24       POTENTIAL.
FT   CHAIN        25   1146       CELL WALL PROTEIN DAN4.
FT   PROPEP     1147   1161       REMOVED IN MATURE FORM (POTENTIAL).
FT   LIPID      1146   1146       GPI-ANCHOR (POTENTIAL).
SQ   SEQUENCE   1161 AA;  118358 MW;  7954C15D69F0CA58 CRC64;
```

alignment_scores:

```
            Quality:  141.00              Length:    436
              Ratio:  0.610                 Gaps:     16
 Percent Similarity:  52.982   Percent Identity: 22.477
```

alignment_block:

US-09-303-518D-465 x DAN4_YEAST ..

Align seg 1/1 to: DAN4_YEAST from: 1 to: 1161

```
   71 CCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGACCGTCAG 120
      |||   |||   :::::::::      :::|||   ::::::::::::
  106 ProAlaIleSerSerAlaLeuSerLysAspGlyIleTyrThrAlaIlePr 122

  121 CATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGGGGAACT 170
         |||      |||::::::::||||||   :::::::::   ::::::
  122 oThrSer......ThrSerThrThrThrThrLysSerSerThrSerThrT 137

  171 TGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGCCATCAGT 220
       |||:::::::::::::   |||      :::|||      :::   :::
  137 hrProThrThrThrIleThrSerThrThrSerThr...ThrSerThrThr 152

  221 TGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTAC 270
         :::|||   :::::::|||      |||             ::::::||
  153 ProThrThrSerThrThrSerThrThrPro.........ThrThrSerTh 166

  271 ATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCT......T 314
       |   |||:::   |||   |||:::|||   |||      |||      |
  166 rThrSerThrThrProThrThrSerThrThrSerThrThrProThrThrS 183

  315 CGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTG 364
       ||||||||                                 :::::::|||
  183 erThrThrSerThrThrProThrThrSerThrThrSerThrThrProThr 199

  365 ACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATCCC 414
       |||:::::::::   |||:::         :::   |||:::|||   ||
  200 ThrSerThrThrSerThrThrProThrThrSerThrThrSerThrThrPr 216

  415 GCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGG 464
       |   |||:::   |||:::         ::::::         |||
  216 oThrThrSerThrThrSerThrThrProThrThrSerThrThrProThr. 232

  465 CGCGAGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCC 514
                    |||:::||||||      :::            :::
```

```
 233 .............ThrSerThrThrSerThrThrSerGlnThrSerThr 244

 515 GCCTCAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGT 564
     ||||||  |||||||||::::::  ::::::       |||::
 245 LysSerThrThrProThrThrSerSerThrSerThr...ThrProThrTh 260

 565 TTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAA 614
     :||||||  |||::::::                ::::|||  ||||:
 260 rSerThrThrProThrThr.................SerThrThrSerT 271

 615 ACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGG............ 652
     ::||||||     ::::::  |||  |||:::
 271 hrAlaProThrThrSerThrThrSerThrThrSerThrThrSerThrIle 287

 653 GCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATC 702
     :::     |||     ||||||:::  :::     |||  :::||
 288 SerThrAlaProThrThrSerThrThrSerSerThrPheSerThrSerSe 304

 703 ATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTAT 752
     |::::::  ::::::  :::|||:::  |||            :::
 304 rAlaSerAlaSerSerValIleSerThrThrAlaThrThrSerThrThrP 321

 753 AAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCA 802
       |||:::     |||           ||||||
 321 heAlaSerLeuThrThrProAlaThrSerThrAla.........SerThr 334

 803 CCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTC 852
         |||         :::||||||       :::     :::::
 335 AspHisThrThrSerSerValSerThrThrAsnAlaPheThrThrSerAl 351

 853 AAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAA 902
     :  |||         :::  :::  |||:::::::|||:::     :
 351 aThrThrThrThrThrSerAspThrTyrIleSerSerSerSerProSerG 368

 903 TGCCGCACAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCC 952
     ::     ::::::  :::|||:::::::::|||        ||||||
 368 lnValThrSerSerAlaGluProThrThrValSerGluValThrSerSer 384

 953 ......CCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGC 996
         |||:::::::::  ::::::                  ::
 385 ValGluProThrArgSerSerGlnVal.....................Th 394

 997 GGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATT 1046
     :::|||  |||:::|||:::     |||::::::  :::  :::
 394 rSerSer..AlaGluProThrThrValSerGluPheThrSerSerValGl 410

1047 GCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACG 1096
      |||      :::|||  |||:::::::::         ::::::  :
 410 uProThrArgSerSerGlnValThrSerSerAlaGluProThrThrValS 427

1097 CCAAATACCCGTCC.........CCTTACCATTCCCGAAATATCCGTTCA 1137
     ::::::::  |||         |||  :::|||  ::::::  |||
 427 erGluPheThrSerSerValGluProThrArgSerSerGlnValThrSer 443

1138 AACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCC 1187
     :::  |||          :::|||  :::||||||||||:::  ||
 444 SerAlaGluProThrThrValSerGlu...PheThrSerSerValGluPr 459

1188 GCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGA 1237
     |           ::::::  ::::::::::::::::  :::|||  |
 459 oThr.........ArgSerSerGlnValThrSerSerAlaGluProThrT 473

1238 CCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTA 1287
     ||         |||  ::::::      |||:::  :::::::::|||
 473 hrValSerGluPheThrSerSerValGluProThrArgSerSerGlnVal 489

1288 AAATACGAT.........ACGAGAATTAATACCGCTGTACCACAAGTGAA 1328
         :::         |||  ::::::      :::  :::|||::
 490 ThrSerSerAlaGluProThrThrValSerGluPheThrSerSerValGl 506

1329 TCCTATA 1335
     :||||||
 506 uProIle 508
```

```
seq_name: SwissProt_40:GTFC_STRMU

seq_documentation_block:
ID   GTFC_STRMU     STANDARD;      PRT;  1375 AA.
AC   P13470; P05427;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   15-DEC-1998 (Rel. 37, Last annotation update)
DE   Glucosyltransferase-SI precursor (EC 2.4.1.5) (GTF-SI)
DE   (Dextransucrase) (Sucrose 6-glucosyltransferase).
GN   GTFC.
OS   Streptococcus mutans.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Streptococcaceae;
OC   Streptococcus.
OX   NCBI_TaxID=1309;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=GS-5;
RX   MEDLINE=89137980; PubMed=2976010;
RA   Ueda S., Shiroza T., Kuramitsu H.K.;
RT   "Sequence analysis of the gtfC gene from Streptococcus mutans GS-5.";
RL   Gene 69:101-109(1988).
RN   [2]
RP   SEQUENCE OF 1-349 FROM N.A.
RC   STRAIN=GS-5;
RX   MEDLINE=87308013; PubMed=3040685;
RA   Shiroza T., Ueda S., Kuramitsu H.K.;
RT   "Sequence analysis of the gtfB gene from Streptococcus mutans.";
RL   J. Bacteriol. 169:4263-4270(1987).
CC   -!- FUNCTION: PRODUCTION OF EXTRACELLULAR GLUCANS, THAT ARE THOUGHT
CC       TO PLAY A KEY ROLE IN THE DEVELOPMENT OF THE DENTAL PLAQUE BECAUSE
CC       OF THEIR ABILITY TO ADHERE TO SMOOTH SURFACES AND MEDIATE THE
CC       AGGREGATION OF BACTERIAL CELLS AND FOOD DEBRIS.
CC   -!- CATALYTIC ACTIVITY: Sucrose + {(1,6)-alpha-D-glucosyl}(N) = D-
CC       fructose + {(1,6)-alpha-D-glucosyl}(N+1).
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- DISEASE: DENTAL CARIES.
CC   -!- MISCELLANEOUS: GTF-I SYNTHESIZES WATER-INSOLUBLE GLUCANS (ALPHA
CC       1,3-LINKED GLUCOSE AND SOME 1,6 LINKAGES), GTF-S SYNTHESIZES
CC       WATER-SOLUBLE GLUCANS (ALPHA 1,6-GLUCOSE). GTF-SI SYNTHESIZES BOTH
CC       FORMS OF GLUCANS.
CC   -!- SIMILARITY: TO OTHER GLUCOSYLTRANSFERASES AND SOME TO A GLUCAN-
CC       BINDING PROTEIN FROM S.MUTANS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M22054; AAA88592.1; -.
DR   EMBL; M17361; AAA88589.1; -.
DR   PIR; JT0345; JT0345.
DR   PIR; C33135; C33135.
DR   InterPro; IPR002479; CW_binding.
DR   InterPro; IPR003318; Glyco_hydro_70.
DR   Pfam; PF01473; CW_binding_1; 7.
DR   Pfam; PF02324; Glyco_hydro_70; 1.
KW   Transferase; Glycosyltransferase; Signal; Repeat; Dental caries.
FT   SIGNAL        1     34
FT   CHAIN        35   1375       GLUCOSYLTRANSFERASE-SI.
FT   DOMAIN       35   1050       CATALYTIC (APPROXIMATE).
FT   DOMAIN     1126   1375       GLUCAN-BINDING (APPROXIMATE).
FT   DOMAIN     1126   1375       2.4 A, 1 C AND 1 AC REPEATS.
FT   REPEAT     1126   1159       A REPEAT.
FT   REPEAT     1169   1200       A REPEAT.
FT   REPEAT     1227   1238       C REPEAT.
FT   REPEAT     1253   1303       AC REPEAT.
FT   REPEAT     1318   1330       A REPEAT (INCOMPLETE).
SQ   SEQUENCE   1375 AA;  153022 MW;  D4B80CBEE0AACE13 CRC64;
```

```
alignment_scores:
           Quality:  140.00              Length:    606
             Ratio:  0.507                Gaps:     32
 Percent Similarity:  45.545   Percent Identity: 22.112

alignment_block:
 US-09-303-518D-465 x GTFC_STRMU   ..

 Align seg 1/1  to: GTFC_STRMU  from: 1  to: 1375

      64 GCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGA 113
         ::::::|||      |||:::|||   |||   :::   :::|||
     643 SerTyrAlaLeuLeuLeuThrAsnLysSerSerValProArgValTyrTy 659

     114 CCGTCAGCATTTCGAACCCGACGGGAAATAC..................... 144
             :::   |||      ||||||:::|||
     659 rGlyAspMetPheThrAspAspGlyGlnTyrMetAlaHisLysThrIleA 676

     145 .........................................CACCTATTC 153
                                                  ::::::
     676 snTyrGluAlaIleGluThrLeuLeuLysAlaArgIleLysTyrValSer 692

     154 GGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCAT.............. 189
         |||:::::::::        ::::::   |||:::
     693 GlyGlyGlnAlaMetArgAsnGlnGlnValGlyAsnSerGluIleIleTh 709

     190 ..........ATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACC 229
                    |||   |||   :::::::::   :::   |||:::
     709 rSerValArgTyrGlyLysGlyAlaLeuLysAlaThrAspThrGlyAspA 726

     230 TGTTCATCCAGCAG.........GCGGCCATTAAAGGAAATATCGGCTAC 270
                 :::            |||:::|||:::||||||
     726 rgThrThrArgThrSerGlyValAlaValIleGluGlyAsnAsnProSer 742

     271 ATT...GTCCGCTTTTCCGAT.........................CA 290
         :::   ::::::   ||||||                        ||
     743 LeuArgLeuLysAlaSerAspArgValValValAsnMetGlyAlaAlaHi 759

     291 CGGGCACGAAGTCCATTCCCCCTTC............GACAAC......C 322
         |   ::::::::::::   |||:::            ||||||
     759 sLysAsnGlnAlaTyrArgProLeuLeuLeuThrThrAspAsnGlyIleL 776

     323 ATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTC 372
           |||   |||||||||:::::   ||||||
     776 ysAlaTyrHisSerAspGlnGluAlaAlaGly.................. 786

     373 AGCCTTTACCGCATCCATTGGGACGGATACGAA......CACCATCCCGC 416
            |||   |||        |||      |||                ||
     787 ...LeuValArgTyrThrAsnAspArgGlyGluLeuIlePheThrAlaAl 802

     417 CGAC......GGCTATGACGGGCCACAGGGCGGCGGCTATCCC....... 453
         ||||      ||||||   :::||||||   :::||||||
     802 aAspIleLysGlyTyrAlaAsnProGlnValSerGlyTyrLeuGlyValT 819

     454 ..GCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGTT 501
           :::|||   ||||||                              ::::::
     819 rpValProValGlyAla..........................AlaAla 826

     502 GCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAACG 551
         |||::::::|||:::      :::          ||||||      ::
     827 AspGlnAspValArgValAlaAlaSerThrAlaProSerThrAspGlyLy 843

     552 GCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAG 601
         :   |||         |||:::::::::::::::   ::::::   :::
     843 sSerVal.........HisGlnAsnAlaAlaLeuAspSerArgValMetP 857

     602 GCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCG 651
           :::||||||::::::         ::::::   :::   ::::::
     857 heGluGlyPheSerAsnPheGlnAlaPheAlaThrLysLysGluGluTyr 873

     652 GGCAATGCCGCCGAAGCTTTCAAC...................GGCAC 680

         |||::::::   |||   |||                      ||||::
     874 ThrAsnValValIleAlaLysAsnValAspLysPheAlaGluTrpGlyVa 890

     681 TGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCG 730
         ::::|||               :::      |||   :::   |||::::
     890 lThrAsp..............PheGluMetAlaProGlnTyrValSerS 902

     731 CAGGCGATGCCGTG.........CAGGGTATAAGCGAAGGCTCAAACATT 771
         ::   |||:::            :::   |||:::::::|||      :::
     902 erThrAspGlySerPheLeuAspSerValIleGlnAsnGlyTyrAlaPhe 918

     772 GCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCG 821
         :::      :::   ||||||   :::|||      ||||||   :::
     919 ThrAspArgTyrAspLeuGly...IleSerLysProAsnLysTyrGlyTh 934

     822 CATCAACGATTTGGCAGAT...ATGGCGCAACTCAAAGACTATGCCGCAG 868
           :::||||||:::      :::      |||   :::   :::
     934 rAlaAspAspLeuValLysAlaIleLysAlaLeuHisSerLysGlyIleL 951

     869 CAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATA 918
           ::::::   ||||||:::   ::::::      |||
     951 ysValMetAlaAspTrpValProAspGlnMetTyrAlaLeuProGluLys 967

     919 GAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGG 968
         |||:::|||:::
     968 GluValValThr..................................... 971

     969 AGCTGTTCGG...GGAAAATACGGCTTGGGCGGCATCACGGCACATCCTG 1015
          |||:::|||      |||||||||                  ||||
     972 .AlaThrArgValAspLysTyrGlyThr.................ProV 982

    1016 TCAAGCGGTCGCAGATG...GGCGAGATCGCATTGCCGAAAGGGAAATCC 1062
         ||      ||||||:::   :::   :::   :::      |||||||||
     982 alAlaGlySerGlnIleLysAsnThrLeuTyrValValAspGlyLysSer 998

    1063 GCCGTCAGCGACAAT.....................TTTGCCGATGCGGC 1091
         :::   :::|||:::                    |||   :::
     999 SerGlyLysAspGlnGlnAlaLysTyrGlyGlyAlaPheLeuGluGluLe 1015

    1092 ATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACT 1141
              ||||||||||||:::   :::      ::::::|||
    1015 uGlnAlaLysTyrProGluLeuPheAlaArgLysGlnIle.......... 1028

    1142 TGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCCGCCG 1191
                                          |||:::   ||||||
    1029 ..............................SerThrGlyValProMet 1034

    1192 TCAAACGGAAAGAATGTGAAA...CTGGCAAACAAACGCCACCCGAAGAC 1238
         :::   :::   ::::::|||      :::   |||            ||
    1035 AspProSerValLysIleLysGlnTrpSerAlaLysTyrPheAsnGlyTh 1051

    1239 CAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAA 1288
         |::::::         |||   |||:::                  |||
    1051 rAsnIleLeuGlyArgGlyAlaGlyTyr.................ValL 1062

    1289 AATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGAT 1338
              |||         ||||||              :::   ::::::
    1062 euLysAspGlnAlaThrAsnThr.........TyrPheSerLeuValSer 1075

    1339 GAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCAT......TC 1382
         :::   :::|||   ||||||:::   |||:::      |||      ||
    1076 AspAsnThrPheLeuProLysSerLeuValAsnProAsnHisGlyThrSe 1092

    1383 TTGGTCTATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTA 1432
         |   |||:::|||:::   :::      :::|||           |||
    1092 rSerServalThrGlyLeuValPheAspGlyLys............GlyT 1105

    1433 GAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCA 1482
           :::   |||            |||   :::   :::|||   :::
    1105 yrValTyrTyrSerThrSerGlyAsnGlnAlaLysAsnAlaPheIleSer 1121

    1483 AAAGGACCTAATAATGGATATTTGGATAAATTTGGTAATGAATGGACTAA 1532
            |||   |||      |||:::|||:::   |||         |||
```

```
    1122 LeuGlyAsnAsnTrpTyrTyrPheAspAsnAsnGlyTyrMetValThr.. 1137

    1533 AGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATGG.........GATG 1573
         |||    :::     :::|||    ::::::    :::          :
    1138 .GlyAlaGlnSerIleAsnGlyAlaAsnTyrTyrPheLeuSerAsnGlyI 1154

    1574 TTCAATTGTCTAAA............ACAGGAAGAGAGCAACTTGGATGG 1611
         ::|||||   :::            :::|||::::::   |||:::::
    1154 leGlnLeuArgAsnAlaIleTyrAspAsnGlyAsnLysValLeuSerTyr 1170

    1612 GCTAGTAGGGATGGTAAG 1629
           ::::::||||||:::
    1171 TyrGlyAsnAspGlyArg 1176

seq_name: SwissProt_40:DRPL_RAT

seq_documentation_block:
ID   DRPL_RAT       STANDARD;      PRT;  1183 AA.
AC   P54258;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   30-MAY-2000 (Rel. 39, Last annotation update)
DE   Atrophin-1 (Dentatorubral-pallidoluysian atrophy protein).
GN   DRPLA.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Cerebellum, and Striatum;
RX   MEDLINE=97317138; PubMed=9173996;
RA   Loev S.J., Margolis R.L., Young W.S., Li S.-H., Schilling G.,
RA   Ashworth R.G., Ross C.A.;
RT   "Cloning and expression of the rat atrophin-I (DRPLA disease gene)
RT   homologue.";
RL   Neurobiol. Dis. 2:129-138(1995).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain, Cerebellum, Hippocampus, and Substantia nigra;
RX   MEDLINE=96081227; PubMed=8541849;
RA   Schmitt I., Epplen J.T., Riess O.;
RT   "Predominant neuronal expression of the gene responsible for
RT   dentatorubral-pallidoluysian atrophy (DRPLA) in rat.";
RL   Hum. Mol. Genet. 4:1619-1624(1995).
CC   -!- TISSUE SPECIFICITY: PREDOMINANT NEURONAL EXPRESSION, ALTHOUGH
CC       MARKEDLY REDUCED AMOUNTS ARE FOUND IN MOST OTHER TISSUES.
CC   -!- DEVELOPMENTAL STAGE: SIMILAR EXPRESSION AT ALL DEVELOPMENT STAGES
CC       (DAY 14.5 P.C., 17.5 P.C., NEWBORNS AND ADULTS).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U31777; AAA80337.1; -.
DR   EMBL; X89453; CAA61623.1; -.
DR   InterPro; IPR002951; Atrophin.
DR   PRINTS; PR01222; ATROPHIN.
FT   DOMAIN      165    171       POLY-PRO.
FT   DOMAIN      303    306       POLY-PRO.
FT   DOMAIN      377    383       POLY-SER.
FT   DOMAIN      387    391       POLY-SER.
FT   DOMAIN      440    446       POLY-PRO.
FT   DOMAIN      477    480       POLY-HIS.
FT   DOMAIN      481    489       POLY-GLN.
FT   DOMAIN      502    505       POLY-PRO.
FT   DOMAIN      562    572       POLY-SER.
FT   DOMAIN      702    705       POLY-PRO.
FT   CONFLICT    455    455       N -> S (IN REF. 2).
FT   CONFLICT    594    594       F -> L (IN REF. 2).
FT   CONFLICT    689    689       P -> R (IN REF. 2).
FT   CONFLICT    717    717       T -> M (IN REF. 2).
FT   CONFLICT    737    737       A -> V (IN REF. 2).
FT   CONFLICT    965    965       MISSING (IN REF. 2).
SQ   SEQUENCE   1183 AA;  124778 MW;  7FB9928DCADF9B1F CRC64;


alignment_scores:
            Quality:  139.00              Length:    595
              Ratio:  0.543                Gaps:     32
 Percent Similarity:  43.025   Percent Identity: 22.017

alignment_block:
 US-09-303-518D-465 x DRPL_RAT   ..

 Align seg 1/1  to: DRPL_RAT  from: 1  to: 1183

     12 CCGCAAAATATCCCTTATTCTGTCCATACTGGC.............AG 46
        |||::::::||||||   :::   ::::::|||             ||
    171 ProAspSerIleProArgGlnProGluSerGlyPheGluProHisProSe 187

     47 TGTGCCTGCCGATGCA.....................TGCACACGCC 72
        |||||||      :::                        :::||||
    187 rValProProThrGlyTyrHisAlaProMetGluProProThrSerArgL 204

     73 TCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGACCG..... 116
        ||     |||               |||   |||:::   |||
    204 euPheGlnGlyPro...........ProProGlyAlaProProProHis 216

    117 ..................TCAGCATTTCGAACCCGACGGGAAATACCACC 148
                           ||||||              :::   |||||
    217 ProGlnLeuTyrProGlySerAlaGlyGlyGlyValLeuSerGlyProPr 233

    149 TATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTG 198
        |:::      :::||||||:::::   ::::::   :::         |
    233 oMetGlyProLysGlyGlyAlaAlaAlaSerSerValGlyProProSerG 250

    199 GG......AAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCA 242
        ||      :::|||      |||::::::     |||:::      :::
    250 lyGlyLysGlnHisProProProThrThrProIleProIleSerSerSer 266

    243 GGCGG.............CCATTAAAGGAAATATCGGCTACATTGTCCGC 279
        |||               |||  |||      ::::::  ::::::||
    267 GlyAlaSerGlyAlaProProAlaLysProProAsnThrProValGlyAl 283

    280 TTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTC 329
        |            :::      |||  ||||||  :::|||:::||||
    283 a.........GlyAsnLeuProSerAlaProProProAlaThrPheProH 297

    330 ACATT.................CCGATTCTGATGAAGCCGGTAGTCCCG 361
        ||:::                |||      ::::::|||:::
    297 isValThrProAsnLeuProProProProAlaLeuArgProLeu...... 311

    362 TTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCAT 411
              ::::::|||   ::::::      |||      ::::::   ||
    312 ......AsnAsnAlaSerAlaSerProProGlyMetGlyAlaGlnProIl 326

    412 CCCG..............CCGACGGCTATGACGGGCCACAGGGCGGCGG 446
        ||||              |||   ||||||   |||:::      ::::
    326 eProGlyHisLeuProSerProHisAlaMet...GlyGlnGlyMetSerG 342

    447 CTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAG 496
        :::::|||           :::|||         |||   :::
    342 lyLeuProProGlyProGluLysGlyProThrLeuAlaProSerProHis 358

    497 GCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGA 546
           ||||||      |||:::|||                  ||||||
    359 ProLeuProProAlaSerSerSer.................AlaProGl 369

    547 CAACGGCTTG............TCGACCGTTTCCACAATACCGGTAGTAT 584
              :::             |||:::   |||:::      |||::::
    369 yProProMetArgTyrProTyrSerSerCysSerSerSerSerValAlaA 386
```

```
 585 GCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCA...CCCGATACA 631
     ::      ::::::   |||:::   |||:::   |||      :::|||
 386 laSerSerSerSerSerAlaAlaThrSerGlnTyrProAlaSerGlnThr 402

 632 GCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACT 681
        ||||||::                ||||||         |||:::::
 403 LeuProSerTyrProHisSerPheProProPro.....ThrSerMe 416

 682 GCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGC 731
     :::::::|||::::::      :::      |||   :::|||   :::|
 416 tSerValSerAsnGlnProProLysTyrThrGlnProSerLeuProSerG 433

 732 AGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGC 781
     |||||:::
 433 lnAlaVal..................................... 435

 782 ACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGAT 831
         |||         ||||||
 436 ......TrpSerGlnGlyProProProPro.....ProProProTy 447

 832 TTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAG...CCATCCG 878
         ::::::      |||::::::|||   |||          |||   :
 447 rGlyArgLeuLeuProAsnAsnAsnThrHisProGlyProPheProProT 464

 879 CGATTGGGCAGTCC............AAAACCCCAATGCCGCACAAGGCA 916
     ::   |||||||||               :::|||      ||| :::||
 464 hrGlyGlyGlnSerThrAlaHisProProAlaProAlaHisHis.HisHi 480

 917 TAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATT 966
     |::::::::::||||||              :::|||:::|||:::
 480 sGlnGlnGlnGlnGlnProGlnProGlnProGlnGlnHisHisH 497

 967 GGAGCTGTTCGGGGAAAATACGGCTTGGG..................... 995
       :::   ||||||           |||
 497 isGlyAsnSerGlyProProProProGlyAlaTyrProHisProLeuGlu 513

 996 .........CGGCATCACGGCACATCCTGTCAAGCGGTCG...CAGATGG 1033
                ||||||  |||||||||   :::   |||   :::::::|
 514 SerSerAsnSerHisHis.AlaHisProTyrAsnMetSerProSerLeuG 530

1034 GCGAGATC...GCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATTTT 1080
     ||::::::         |||   |||   :::   :::       :::
 530 lySerLeuArgProTyrProProGlyProAlaHisLeuProProSerHis 546

1081 GCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATAT 1130
     ::::::::::::|||:::
 547 GlyGlnValSerTyrSer............................... 552

1131 CCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAA 1180
                       :::   |||   ::::::         |||:
 553 ...................GlnAlaGlyProAsnGlyProProValSerS 563

1181 CCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCA...AACAAACGC 1227
     ::         |||:::|||::::::   :::   |||   :::
 563 erSerSerAsnSerSerGlySerSerSerGlnAlaAlaTyrSerCysSer 579

1228 CACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGA 1277
     ||||||:::::::::         :::|||   ::::::|||
 580 HisProSerSerSerGlnGlyProGlnGlyAlaSerTyrPro........ 593

1278 AAAAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGA 1327
                                          |||   |||
 594 ....................................PheProProValP 598

1328 ATCCTATAGATGAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCT 1377
       ||||||                          :::::::::||||||
 598 roProIle.........................ThrThrSerSerAla 605

1378 CATTCTTGGTCTATAACTGCCAGAATT..................... 1404
             ::::::   |||   :::
 606 ThrLeuSerThrValIleAlaThrValAlaSerSerProAlaGlyTyrLy 622

1405 .......................CAATACGCAAAA...TTACCAAGGC 1426
                              ||||||:::|||      |||
 622 sThrAlaSerProProGlyProProGlnTyrSerLysArgAlaProSerP 639

1427 AAGGTAGAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCG 1476
       |||      :::         |||   :::|||:::|||::::::|||
 639 roGlySerTyrLysThrAlaThrProProGlyTyrLysProGlySerPro 655

1477 ......CTACCAAAAGGACCTAATAATGGATATTTGGATAAATTTGGTAA 1520
           :::   |||         ||||||           ||||::
 656 ProSerPheArgThrGlyThrProProGlyTyrArg.........GlyTh 669

1521 TGAATGGACTAAAGGTCCATCAAGAACTAAA 1551
     ::::         ||||||:::      |||
 669 rSerProProAlaGlyProGlyThrPheLys 679
```

```
seq_name: SwissProt_40:VRP1_YEAST

seq_documentation_block:
ID   VRP1_YEAST     STANDARD;      PRT;   817 AA.
AC   P37370; Q06133;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   01-NOV-1997 (Rel. 35, Last annotation update)
DE   Verprolin.
GN   VRP1 OR MDP2 OR END5 OR YLR337W OR L8300.13.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=A364;
RX   MEDLINE=95058201; PubMed=7968536;
RA   Donnelly S.F.H., Pocklington M.J., Pallota D., Orr E.;
RT   "A proline-rich protein, verprolin, involved in cytoskeletal
RT   organization and cellular growth in the yeast Saccharomyces
RT   cerevisiae.";
RL   Mol. Microbiol. 10:585-596(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Johnston M., Andrews S., Brinkman R., Cooper J., Ding H., Du Z.,
RA   Favello A., Fulton L., Gattung S., Greco T., Kirsten J.,
RA   Kucaba T., Hallsworth K., Hawkins J., Hillier L., Jier M.,
RA   Johnson D., Johnston L., Langston Y., Latreille P., Le T.,
RA   Mardis E., Menezes S., Miller N., Nhan M., Pauley A., Peluso D.,
RA   Rifken L., Riles L., Taich A., Trevaskis E., Vignati D.,
RA   Wilcox L., Wohldman P., Vaudin M., Wilson R., Waterston R.;
RL   Submitted (JAN-1995) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: INVOLVED IN CYTOSKELETAL ORGANIZATION AND CELLULAR
CC       GROWTH. MAY EXERT ITS EFFECTS ON THE CYTOSKELETON DIRECTLY, OR
CC       INDIRECTLY VIA PROLINE-BINDING PROTEINS (E.G. PROFILIN) OR
CC       PROTEINS POSSESSING SH3 DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z26645; CAA81388.1; -.
DR   EMBL; U19028; AAB67263.1; -.
DR   PIR; S39626; S39626.
DR   SGD; S0004329; VRP1.
DR   InterPro; IPR003124; WH2.
DR   Pfam; PF02205; WH2; 2.
DR   SMART; SM00246; WH2; 2.
KW   Cytoskeleton; Repeat.
FT   DOMAIN        5     14       POLY-PRO.
FT   DOMAIN      239    245       POLY-PRO.
```

```
FT   DOMAIN       349    357       POLY-PRO.
FT   DOMAIN       396    406       POLY-PRO.
FT   DOMAIN       424    431       POLY-PRO.
FT   DOMAIN       462    468       POLY-SER.
FT   DOMAIN       704    708       POLY-PRO.
FT   CONFLICT     308    308       P -> R (IN REF. 1).
FT   CONFLICT     350    350       A -> R (IN REF. 1).
FT   CONFLICT     689    689       V -> E (IN REF. 1).
FT   CONFLICT     710    817       PSTMDTGTSNSPSKNLKQRLFSTGGSTLQHKHNTHTNQPDV
FT                                 DVGRYTIGGSNSIVGAKSGNERIVIDDSRFKWTNVSQMPKP
FT                                 RPFQNKTKLYPSGKGSSVPLDLTLFT -> HLRWIPVPLIA
FT                                 PVKTLNNGYFLQVDRRCNTSIIRIQINQMLM (IN REF.
FT                                 1).
SQ   SEQUENCE   817 AA;  82593 MW;  24C7522D5B1CA1C8 CRC64;


alignment_scores:
           Quality:  136.50            Length:    502
             Ratio:  0.626               Gaps:     27
 Percent Similarity:  43.426   Percent Identity: 24.502

alignment_block:
 US-09-303-518D-465 x VRP1_YEAST   ..

 Align seg 1/1  to: VRP1_YEAST  from: 1  to: 817

     155 GCAGCAGGGGGGAACTTGCCGAGCGCA..................... 181
         |||   ||||||:::   |||:::
      15 AlaLeuGlyGlySerAlaProLysProAlaLysSerValMetGlnGlyAr 31

     182 ....GCGGTCATATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAA 227
             |||:::::::::|||   :::           ::::::   |||
      31 gAspAlaLeuLeuGlyAspIleArgLysGlyMetLysLeuLysLysAlaG 48

     228 CCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATTGTCC 277
             ::::::         |||:::         :::::::::   |||
      48 luThrAsnAspArgSerAlaProIleValGlyGlyGlyValValSerSer 64

     278 GCTTTTCCGATCACGGGCACG...AAGTCCATTCCCCCTTCGACAACCAT 324
         |||         :::||||||   :::|||      ||||||   :::
      65 AlaSerGlySerSerGlyThrValSerSerLysGlyProSerMetSerAl 81

     325 GCCTCACATTCCGATTCTGATGAAGCCGGTA..................... 355
          |||   ||||||   :::      |||
      81 aProProIleProGlyMetGlyAlaProGlnLeuGlyAspIleLeuAlaG 98

     356 .....GTCCCG............TTGACGGATTCAGCCTTTACCGCATCC 388
             :::|||            ::::::::::::|||   |||
      98 lyGlyIleProLysLeuLysHisIleAsnAsnAsnAlaSerThrLysPro 114

     389 ATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAG 438
             :::   :::   :::   ||||||        :::
     115 ......SerProSerAlaSerAlaProProIleProGlyAlaValProSe 129

     439 GGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGA 488
          :::|||||||   |||:::|||:::|||                ::::
     129 rValAlaAlaProProIleProAsnAla............ProLeuSerP 142

     489 CATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCA 538
                 |||:::|||:::|||   :::|||:::   |||
     142 roAlaProAlaValProSerIleProSerSerSerAlaPro......Pro 156

     539 GCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTG 588
           ||||||         |||::::::      ||||||:::|||
     157 IleProAspIleProSerSerAlaAlaProProIleProIleVal..... 171

     589 ACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCC............. 625
                            :::|||   |||||||||
     172 ................ProSerSerProAlaProProLeuProLeuSerG 183

     626 ..GATACAGCCCCGAGCTGG...ACAGATCGGGCAATGCCGCCGAAGCTT 670
             :::||||||:::         :::|||         |||::::::
     183 lyAlaSerAlaProLysValProGlnAsnArgProHisMetProSerVal 199

     671 TCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAA 720
             |||                              ::::::|||:::||
     200 ArgProAla.....................HisArgSerHisGlnArgLy 209

     721 ATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACAT 770
         |   |||      :::::::|||
     209 sSerSerAsnIleSerLeuProSer........................ 217

     771 TGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGC 820
                              |||      ||||||
     218 ...................ValSerAlaProProLeuPro......... 224

     821 GCATCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCA 870
                                      |||   :::::::|||   ::
     225 ............................SerAlaSerLeuProThrHi 231

     871 GCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCG............. 907
         :   |||        |||:::      |||   |||
     231 sValSerAsnProProGlnAlaProProProProThrProThrIleG 248

     908 ........CACAAGGCATAGAAGCCG......TCAGCAATATCTTTACGG 943
                 |||      ||||||      :::|||:::|||
     248 lyLeuAspSerLysAsnIleLysProThrAspAsnAlaValSerProPro 264

     944 CAGTCATCC...CCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGC 990
         :::|||:::   |||:::   ||||||   :::::::::|||   :::||
     265 SerSerGluValProAlaGlyGlyLeuProPheLeuAlaGluIleAsnAl 281

     991 TTGGGCGGCATCACGGCACATC.........CTGTCAAGCGGTCGCAGAT 1031
         |      ::::::|||                :::||||||   :::
     281 aArgArgSerGluArgGlyAlaValGluGlyValSerSerThrLysIleG 298

    1032 GGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCG.........TCAGCG 1072
           ::::::::::|||         ::::::||||||         ||||||
     298 lnThrGluAsnHisLysSerProSerGlnProProLeuProSerSerAla 314

    1073 ACAATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCC 1122
               :::|||      |||:::|||         |||   :::   ||
     315 ProProIleProThrSerHisAlaProProLeuProProThrAlaProPr 331

    1123 CGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCA. 1171
         |      |||:::   :::   :::|||:::      |||   ||||||
     331 oProProSerLeuProAsnValThrSerAlaProLysLysAlaThrSerA 348

    1172 ................CCTCCTCAACCGTGCCGCCGTCAAACGGAAAGA 1204
                         ||||||   ||| ||||||:::::      ::
     348 laProAlaProProProProProLeuPro.AlaAlaMetSerSerAlaSe 364

    1205 ATGTGAAACTGGCAAACAAACGCCA...CCCGAAGACCAAAGTGCCGTTT 1251
         :   ::::::   |||   ||||||   |||    |||   :::|||
     364 rThrAsnSerValLysAlaThrProValPro.ProThrLeuAlaPro... 379

    1252 GACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAG 1301
                  :::|||||||
     380 .........ProLeuProAsn............................ 383

    1302 AATTAATACCGCTGTACCA............CAAGTGAATCCTATAGATG 1339
             :::|||:::||||||            :::      |||
     384 ....ThrThrSerValProProAsnLysAlaSerSerMetProAlaProP 399

    1340 AACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCT 1389
           |||         |||      ::::::   |||:::   |||
     399 roProProProProProProProGlyAlaPheSerThrSerSerAlaLeu 415

    1390 ATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAG 1439
           ::::::   |||      |||   ||||||
     416 SerAlaSerSerIleProLeuAlaProLeuProProPro........... 428

    1440 ATATATCCCACCTAAAAATTACTCT.........CCTTCAGCACCGCTAC 1480
                 ||||||   :::   :::         ||||||||||||   |
     429 .......ProProProSerValAlaThrSerValProSerAlaProProP 443
```

```
   1481 CA 1482
        ||
    443 ro 443

seq_name: SwissProt_40:YGY3_HALSQ

seq_documentation_block:
ID   YGY3_HALSQ     STANDARD;      PRT;   437 AA.
AC   P21561;
DT   01-MAY-1991 (Rel. 18, Created)
DT   01-MAY-1991 (Rel. 18, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Hypothetical 50.6 kDa protein in the 5'region of GYRA and GYRB (ORF
DE   3).
OS   Haloferax sp. (strain Aa 2.2).
OC   Archaea; Euryarchaeota; Halobacteriales; Halobacteriaceae; Haloferax.
OX   NCBI_TaxID=2254;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91100352; PubMed=1846146;
RA   Holmes M.L., Dyall-Smith M.L.;
RT   "Mutations in DNA gyrase result in novobiocin resistance in
RT   halophilic archaebacteria.";
RL   J. Bacteriol. 173:642-648(1991).
CC   ------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ------------------------------------------------------------------------
DR   EMBL; M38373; -; NOT_ANNOTATED_CDS.
DR   PIR; C39135; C39135.
KW   Hypothetical protein.
SQ   SEQUENCE   437 AA;  50626 MW;  B5B99A2AF3892BEF CRC64;


alignment_scores:
           Quality:  135.00            Length:    466
             Ratio:  0.689              Gaps:     27
 Percent Similarity:  42.060   Percent Identity: 25.107

alignment_block:
 US-09-303-518D-465 x YGY3_HALSQ   ..

 Align seg 1/1  to: YGY3_HALSQ  from: 1  to: 437

     276 CCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATG 325
         |||      |||   |||   |||:::      ||||||
      21 ProArgThrArgArgArgHisArgAsnAspHisProLeu.........Le 34

     326 CCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGC 375
                 |||:::         |||   :::|||   |||:::|||
      34 uAlaGlyArgArgTyr......LeuArgAspAspArgValArgLeuGlnA 49

     376 CTTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTA 425
                 |||   :::|||   |||:::|||:::   |||
      49 spAlaArgLysProProAlaArgAlaArgValProGlnLeuArg...... 63

     426 TGACGGGCCACAGGGCGG......CGGCTATCCCGCTCCCAAAGGCGCGA 469
                   ||||||         |||   |||::::::||||||
      64 ............GlyArgAspPheAlaLeuArgArgAlaAspArgArgVa 76

     470 GGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTC 519
             ::::::   ||||||   ::::::         :::      ||||
      76 lGluHisValProLeuArgGlyArgHisProArgValArgArgValProG 93

     520 AACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCA 569
         ||   |||:::
      93 lnArgAspGln..................................... 96

     570 CAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCG 619
                         |||:::      ||||||            ||
      97 .................AspGlyAlaProArgArg............Ar 103

     620 CCACCCGATACAGCCCCGA...GCTGGACAGATCGGGCAATGCCGCCGAA 666
         ||||   ::::::   |||   :::|||:::   |||:::   |||:::
     103 gHisLeuLeuArgArgArgValGlyGlyHisArgGlyArgAsnArgHisA 120

     667 GCTTTCAACGGCACTGC..............AGATATCGTCAAAAACAT 701
         ::   :::|||::::::                |||   ||||||:::|||
     120 laGlyAspArgArgAlaProGlyValAspSerArgLeuArgGlnGlnHis 136

     702 ...CATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGG 748
            |||   |||||||||:::
     137 GlnHisProArgGlyArgHis........................... 143

     749 GTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTT 798
                                                        |||
     144 .............................................AlaS 145

     799 TCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCA 848
             |||   |||||||||||||||   |||      |||
     145 erAspArgValGlnAspGlyAlaHisProArgArgGlnArgLeuArgGlu 161

     849 ACTCAAAGACTATGC............CGCAGCAGCCATCCGCGATTGGG 886
             |||   :::             |||   :::   ||||||   ||
     162 GlnProArgHisAlaGlyArgProArgArgArgGlnProProArgArgGl 178

     887 C.................................AGTCCAAAACCCC 900
         |                                 :::|||:::|||
     178 yArgSerArgGlyThrHisArgArgHisLeuArgGlnAlaProArgProA 195

     901 AATGCCGC.................................ACAAGG 914
             |||                                   |||
     195 laValArgGlyProAspGluAspGlnAlaArgGluPheArgGlyProArg 211

     915 CATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGA 964
         ||||||   ::::::                |||  |||      :::
     212 HisArgArgGluArg.................His.ProProThrAlaA 222

     965 TTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCT 1014
             ::::::||||||:::   |||   |||    ::::::|||:::
     222 rgAspValLeuArgGlyGluProGlyHisGlyAsp.GlyHisHisLeu.. 237

    1015 GTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGC 1064
             :::|||||||||| ||||||   |||          ||||||
     238 ....GluGlyArgArg.GlyArgProArgProGlnGlyArgGluAlaGly 252

    1065 .......................CGTCAGCGACAATTTTGCCGATGCGG 1090
                                 |||   |||  :::     :::||
     253 ArgGlyAlaHisProProGlnValArgAlaArgIleTyrLeuAlaAlaGl 269

    1091 C...ATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCA 1137
         |      |||   :::|||   |||   |||:::
     269 yGluAlaArgGlyLeuProGluProArgProLeuGlyValArgThrValH 286

    1138 AACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCC 1187
             |||:::   ||||||   |||      :::         |||
     286 isArgGlyGlyArgLeuArgGlyArgValGlyGlnAlaGlyProArgPro 302

    1188 GCCGTC..............AAACGGAAAGAATGTGAAACTGGCAAACA 1222
         |||                 :::   |||   :::      :::::
     303 GlnValProGlyAspPheAlaProGlnGlyGluAspSerGluArgArgGl 319

    1223 AACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAAT 1272
         :||||||||||:::   :::|||         |||         ::::::
     319 uThrProProArgProHisSerArgLysArgArgAspThrGlyAlaHisH 336

    1273 TT............TGAAAAAGACGTAAAATACGATACGAGAATTAATAC 1310
                       :::||||||::::::|||::: |||   :::
     336 isArgHisTrpArgArgArgArgArgArgValArgHis.Arg...GluGl 351

    1311 CGCTGTACCACAAGTGAATCCTATAGATGAACCCGTCTTTAATCCTAAAG 1360
```

```
          |||:::|||   ::::::|||   |||:::         :::   ::::
      351 yAlaLeuProAlaAlaHisProAspAspArgArgArgArgArgArgA 368

     1361 GTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTCAATAC 1410
          ::         :::|||                           |||
      368 laHisProAspAlaAla.........................AlaTyr 375

     1411 GCAAAATTACCA..............AGGCAAGGTAGAATCAGATATAT 1445
          |||::::::|||              ::::::||||||:::|||   ::
      376 AlaSerValProAlaHisAlaProAlaHisArgGlyArgLeuArg...Va 391

     1446 CCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAA 1485
          :      ::::::   :::         ||||||||||:::
      391 lArgGlySerThrAlaAlaValProArgProLeuProArg 404

seq_name: SwissProt_40:MUC1_MESAU

seq_documentation_block:
ID   MUC1_MESAU     STANDARD;      PRT;   676 AA.
AC   Q60528;
DT   15-JUL-1999 (Rel. 38, Created)
DT   15-JUL-1999 (Rel. 38, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Mucin 1 precursor.
GN   MUC1.
OS   Mesocricetus auratus (Golden hamster).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Cricetinae;
OC   Mesocricetus.
OX   NCBI_TaxID=10036;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Tracheal epithelium;
RX   MEDLINE=96326118; PubMed=8703480;
RA   Park H., Hyun S.W., Kim K.C.;
RT   "Expression of MUC1 mucin gene by hamster tracheal surface epithelial
RT   cells in primary culture.";
RL   Am. J. Respir. Cell Mol. Biol. 15:237-244(1996).
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON (BY SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: Type I membrane protein.
CC   -!- PTM: HIGHLY O-GLYCOSYLATED AND PROBABLY ALSO N-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; U36918; AAB53965.1; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat.
FT   SIGNAL        1     25       POTENTIAL.
FT   CHAIN        26    676       MUCIN 1.
FT   DOMAIN       26    582       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    583    603       POTENTIAL.
FT   DOMAIN      604    676       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN      458    573       SEA.
FT   CARBOHYD    291    291       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    323    323       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    350    350       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    380    380       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    400    400       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    413    413       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    435    435       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    479    479       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    496    496       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    536    536       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   676 AA;  67616 MW;  95F479B6EC5C3884 CRC64;


alignment_scores:
           Quality:  132.50            Length:    570
             Ratio:  0.496               Gaps:     27
 Percent Similarity:  46.842   Percent Identity: 22.632

alignment_block:
 US-09-303-518D-465 x MUC1_MESAU   ..

 Align seg 1/1  to: MUC1_MESAU  from: 1  to: 676

      55 CCGATGCATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCA 104
         |||:::|||:::   :::|||      |||::::::|||   :::   ::
      40 ProValHisSerGlySerSerAlaProAlaThrSerSerAlaValAspSe 56

     105 GGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCG 154
         ::::            ||| |||:::   ::::::      |||   |||
      56 rAlaThrThrProGlyHis.SerGlySerSerAlaProProThrSerSer 72

     155 GCAGCA...GGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGA 201
         |||:::   :::::::::   |||:::   :::      |||
      73 AlaValAsnSerAlaThrThrProGlyHisSerGlySerSerAlaProPr 89

     202 AACATACAAAGCCATCAGTTGGGCAACCTGT.........TCATCCAGCA 242
          |||   :::|||::::::   ||||||            |||:::|||
      89 oThrSerSerAlaValAsnSerAlaThrThrProValHisSerGlySerS 106

     243 GGCGGCCATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACG 292
           |||:::   :::   :::::::::   ::::::   |||:::
     106 erAlaProValThrSerSerAlaValAsnSerAlaThrThrProValHis 122

     293 GGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTCAC.........AT 333
         :::   :::|||   ||||||:::::::::      :::
     123 SerGlySerSerAlaProProThrSerSerAlaValAsnSerAlaThrTh 139

     334 TCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCG 383
          |||:::                :::|||:::|||:::||||||   ::::
     139 rProValHisSerGlySerSerAlaProValThrSerSerAlaValAsnS 156

     384 CA.........................................TCCATT 391
         ::                                          |||
     156 erAlaThrThrProValHisSerGlySerSerAlaProValThrSerSer 172

     392 GGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGC 441
         ::::::|||:::   |||   |||                    |||   ::
     173 AlaValAspSerAlaThrThrProVal.............HisSerGl 184

     442 GGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACAT 491
         :::::::   |||   |||   :::   ::::::   :::||||||||||
     184 ySerSerAlaPro...ProThrSerSerAlaValAsnSerAlaThrThr. 199

     492 AAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCA 541
                   |||      |||:::|||:::   |||:::|||::::::
     200 ...........ProValHisSerGlySerSerAlaProValThrSerSer 212

     542 CCGGACAACGGCTTGTCGACCGTT............TCCACAATACCGGT 579
               |||:::   :::|||                |||:::   |||||
     213 AlaValAsnSerAlaThrThrProValHisSerGlySerSerAlaProVa 229

     580 AGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATA 629
         |                        |||:::|||   :::         |
     229 l.....................ThrSerSerAlaValAsnSerAlaT 238

     630 CAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCA 679
         ||:::|||      :::      :::   ||||||
     238 hrThrProValHisSerGlySerSerAlaProPro.............. 249

     680 CTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGC 729
                 ||||||   ::::::||||||               |||::
```

```
 250 ......ThrSerSerValValAsnSerAlaThrThrProValHisSerGl 264

 730 GCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTAT 779
     :::::::   |||                  ::::::::::::   |||:
 264 ySerSerAlaPro.............ProThrSerSerAlaValAsnLeuA 277

 780 GCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACG 829
     ::|||:::   |||              :::|||             ||||||
 277 laThrThrProValHisSerGlySerSerThrProAlaThrAsnSerThr 293

 830 ATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGC 879
                       :::|||   |||   |||      |||:::::
 294 ..............ThrAspSerAlaThrThrProValProProGlySe 305

 880 G.......................ATTGGGCAGTCCAAAACCCCAATGC 905
     :                          |||::::::::::||||||:::
 305 rSerMetGlnThrThrGluAlaIleSerGlySerAlaAsnThrProIleH 322

 906 CGCACAAGGCATAGAAGCCG.........TCAGCAATATCTTTACGGCAG 946
       :::   :::       |||         ||||||:::
 322 isAsnGlySerLeuValProThrThrSerSerAlaLeuValProThrThr 338

 947 TCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGC 996
     |||:::      :::|||
 339 SerAlaAlaHisSerGly.............................. 344

 997 GGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATT 1046
      ||||||           ||||||            |||   :::
 345 .AlaSerAlaMetThrAsnSerSerGluSerAspLeuAlaThrThrProI 361

1047 GCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACG 1096
          :::|||:::      |||:::|||                 :::
 361 leAspSerGlyThrSerIleSerThrThr.............LysAla 372

1097 CCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAG 1146
     |||   |||   |||:::       ::::::   |||   |||   ||
 373 ProAlaThrThrProValHisAsnGlySerLeu...ValProThrThrSe 388

1147 CAGCGTTACGGCAAAGAAAACATCACC..................... 1173
     |||||||   ::::::   |||:::
 388 rSerValLeuGlySerAlaThrThrLeuIleHisAsnAspThrSerThrM 405

1174 ..TCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAAC 1221
       ::::::|||:::|||   :::||||||   :::   ::::::   :::
 405 etAlaThrThrThrProValGlyAsnGlyThrGlnSerSerValProSer 421

1222 AAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAA 1271
        |||||||||   |||   :::|||          ::::::     ::
 422 ...ArgHisProValThrProThrProProAlaValSerSerAsnSerTh 437

1272 TTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACC.......... 1311
     ::::      ::::::   |||:::|||   ::::::
 437 rIleAlaLeuSerThrTyrTyrSerThrAlaLeuSerProAlaPheSerS 454

1312 .....GCTGTACCACAAGTG............................. 1326
          |||:::|||||||||
 454 erHisAlaAlaProGlnValSerValGlyValSerPhePheLeuLeuSer 470

1327 ..........................AATCCTATAGATGAACCCGTCT. 1348
                               :::   ::::::::: ||||||
 471 PheHisIleTrpAsnHisGlnPheAsnSerSerLeuGluAsp.ProSerS 487

1349 ....................................TTAATCCTAAAG 1360
                                         |||:::|||:::
 487 erAsnTyrTyrGlnGluLeuLysArgAsnValSerGlyLeuPheLeuGln 503

1361 GTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTCAATAC 1410
     |||:::|||   |||   :::||||||:::         ||||||:::::
 504 ValPheSer...ArgAlaPheLeuGlyIleSerThrIleGluPheArgSe 519

1411 GCAAAATTAC.........................CAAGGCAAGGTAG 1433
     :   :::                           :::|||   |||:
 519 rGlySerValValValAspSerThrValIlePheArgGluGlyAlaValA 536

1434 AATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAA 1483
     :::::      |||::::::|||      ||||||||||||:::   :::
 536 snAla......SerGluValLysSerGlnLeuLeuGlnHisGluGlnGlu 550

1484 AAGGAC 1489
        :::
 551 AlaGlu 552
```

```
seq_name: SwissProt_40:VGP3_EBVA8

seq_documentation_block:
ID   VGP3_EBVA8     STANDARD;      PRT;   886 AA.
AC   Q07284;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   01-OCT-1996 (Rel. 34, Last annotation update)
DE   Envelope glycoprotein GP340 (Membrane antigen) (MA).
GN   BLLF1.
OS   Epstein-barr virus (strain AG876) (Human herpesvirus 4), and
OS   Epstein-barr virus (strain P3HR-1) (Human herpesvirus 4).
OC   Viruses; dsDNA viruses, no RNA stage; Herpesviridae;
OC   Gammaherpesvirinae; Lymphocryptovirus.
OX   NCBI_TaxID=82830, 82829;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=AG876, AND P3HR-1;
RX   MEDLINE=93331716; PubMed=8393237;
RA   Lees J.F., Arrand J.E., Pepper S.V., Stewart J.P., Mackett M.,
RA   Arrand J.R.;
RT   "The Epstein-Barr virus candidate vaccine antigen gp340/220 is highly
RT   conserved between virus types A and B.";
RL   Virology 195:578-586(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=P3HR-1;
RA   Klein K., Mueller-Lantzsch N.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: RESPONSIBLE FOR EBV BINDING TO THE CR2 RECEPTOR ON HUMAN
CC       B-CELLS.
CC   -!- SUBCELLULAR LOCATION: MOST ABUNDANT COMPONENT OF THE VIRAL
CC       ENVELOPE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L07922; AAA02783.1; -.
DR   EMBL; L07923; AAA02787.1; -.
DR   EMBL; X67776; CAA47986.1; -.
KW   Membrane; Glycoprotein; Antigen; Late protein.
FT   CARBOHYD     47     47       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     87     87       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    114    114       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    166    166       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    169    169       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    195    195       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    229    229       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    277    277       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    318    318       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    328    328       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    345    345       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    356    356       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    378    378       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    386    386       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    411    411       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    435    435       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    443    443       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    457    457       N-LINKED (GLCNAC. . .) (POTENTIAL).
```

```
FT   CARBOHYD    497    497       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    519    519       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    533    533       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    568    568       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    589    589       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    624    624       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    680    680       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   886 AA;  92388 MW;  4394F6130DECCA8A CRC64;


alignment_scores:
           Quality:  131.50            Length:    588
             Ratio:  0.502               Gaps:     28
 Percent Similarity:  44.558   Percent Identity: 22.279

alignment_block:
 US-09-303-518D-465 x VGP3_EBVA8   ..

 Align seg 1/1  to: VGP3_EBVA8  from: 1  to: 886

    110 TCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCG.:... 154
        :::|||:::::::|||:::          |||:::|||||||||
    318 AsnThrThrAspIleThrTyrValGlyAspAsnAlaThrTyrSerValPr 334

    155 ..........GCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGG 194
                   :::      :::|||   |||:::::::::|||   :::
    334 oMetValThrSerGluAspAlaAsnSerProAsnValThrValThrAlaP 351

    195 ATTGG..................GAAACATACAAAGCCATCAGTTGGGCA 226
          |||                  ||||||            :::|||:::
    351 heTrpAlaTrpProAsnAsnThrGluThrAspPheLysCysLysTrpThr 367

    227 ACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATTGTC 276
              |||::::::        ::::::||||||:::::::::::
    368 LeuThrSerGlyThrProSerGlyCysGluAsnIleSerGlyAlaPheAl 384

    277 CGCT.........TTTCCGATCACGGGC..................... 295
        ::::         |||   ||||||
    384 aSerAsnArgThrPheAspIleThrValSerGlyLeuGlyThrAlaProL 401

    296 ..............ACGAAGTCCATTCCCCCTTCGACAACCATGCCTCAC 331
                     |||::::::         :::||||||      |||
    401 ysThrLeuIleIleThrArgThrAlaThrAsnAlaThrThrThrThrHis 417

    332 ATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTAC 381
              |||:::   |||                :::|||:::   ||
    418 LysValIlePheSerLys............AlaProGluSerThrThrTh 430

    382 CGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGG 431
        |:::         :::   |||:::         |||::::::   ||||
    430 rSerProThrLeuAsnThrThrGlyPheAlaAlaProAsnThrThrThrG 447

    432 GCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACA 481
        ||                     ||||||::::::      :::   |||
    447 ly....................LeuProSerSerThrHisValProThr 456

    482 GCTACGACATAAAAGGCGTTG......CCCAAAATATCCGCCTCA..... 520
              |||      |||         |||   :::|||::::::
    457 AsnLeuThrAlaProAlaSerThrGlyProThrValSerThrAlaAspVa 473

    521 .ACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCA 569
         |||   ||||||   |||:::      :::||||   |||   |||:::
    473 lThrSerProThrProAlaGlyThrThrSerGlyAlaSerProValThrP 490

    570 CAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCG 619
             |||               ::::::   :::|||:::|||:::|||
    490 roSerProSer.........ProArgAspAsnGlyThrGluSerLysAla 503

    620 CCACCCGATACAGCCCCG...AGCTGGACAGATCGGGCAATGCCGCCGAA 666
        |||      |||:::|||   |||   :::            |||
    504 ProAspMetThrSerProThrSerAlaValThrThrProThrProAsnAl 520

    667 GCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGG 716

        |||   :::      :::::::::   |||   :::|||   :::
    520 aThrSerProThrProAlaValThrThrProThrProAsnAlaThrSerP 537

    717 AGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAA 766
            |||:::      :::            |||   :::   :::
    537 roThrLeuGlyLysThrSerProThrSerAlaValThrThrProThrPro 553

    767 ACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATG 816
        :::            |||      |||      |||      :::
    554 AsnAlaThrSerProThrProAlaValThrThrProThrProAsn..... 568

    817 GCGCGCATCAACGATTTGGCAGATATGGCGCAACTCA............. 853
              :::||||||      :::   :::::::|||
    569 .......AlaThrIleProThrLeuGlyLysThrSerProThrSerAlaV 583

    854 .....................................AAGACTATG 862
                                             :::|||
    583 alThrThrProThrProAsnAlaThrSerProThrValGlyGluThrSer 599

    863 CCGCAGCAGCCATCCGCGATT............GGGCAGTCCAAAACCCC 900
        ||||||      ::::::               |||   |||:::|||||
    600 ProGlnAlaAsnThrThrAsnHisThrLeuGlyGlyThrSerSerThrPr 616

    901 AATG.........CCGCACAAG........................... 913
        |:::       |||   |||
    616 oValValThrSerProProLysAsnAlaThrSerAlaValThrThrGlyG 633

    914 ..............GCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCA 949
                       :::   :::   |||:::::|||||||||   |||
    633 lnHisAsnIleThrSerSerSerThrSerSerMetSerLeuArgProSer 649

    950 TCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGC 999
        |||   |||          |||   ::::::   :::|||:::      ::
    650 SerIleSer.........GluThrLeuSerProSerThrSerAspAsnSe 663

   1000 ATCACGGCAC......ATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGC 1043
        ::::   |||      :::|||:::|||::::::      :::   ::::
    663 rThrSerHisMetProLeuLeuThrSerAlaHisProThrGlyGlyGluA 680

   1044 ATTGCCGAAAGGGAAATCCGCCGTCA...GCGACAATTTTGCCGATGCGG 1090
        ::      :::   :::|||   |||   :::|||
    680 snIleThrGlnValThrProAlaSerThrSerThrHisHisValSerThr 696

   1091 CATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGT...... 1134
           :::|||      ||||||   |||       :::   |||
    697 SerSerProAlaProArgProGlyThrThrSerGlnAlaSerGlyProGl 713

   1135 ....TCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAA 1180
            |||:::   :::   :::   |||:::   |||:::|||::::::|
    713 yAsnSerSerThrSerThrLysProGlyGluValAsnValThrLysGlyT 730

   1181 CCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCAC 1230
        ||   ||||||:::|||:::   :::
    730 hr...ProProLysAsnAlaThrSer....................... 737

   1231 CCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAA 1280
        |||::::::                              |||:::   :::||
    738 ProGlnAla.........................ProSerGlyGlnLy 745

   1281 AGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATC 1330
        |                              ||||||||||||   |||:::
    745 s.........................ThrAlaValProThrValThrS 753

   1331 CTATAGATGAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCAT 1380
                              :::   ::::::::::::|||:::   |||
    753 erThrGlyGly.........LysAlaAsnSerThrThrGlyGlyLysHis 766

   1381 TCTTGGTCTATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGG 1430
        :::   :::      ||||||   :::                    ||
    767 ThrThrGlyHisGlyAlaArgThrSerThrGluProThrThrAspTyrGl 783

   1431 TAGAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCTAC 1480
        |               |||      :::|||:::   ::::::   ||||
```

```
     783 yGlyAspSerThrThrProArgThrArgTyrAsnAlaThrThrTyrLeuP 800

    1481 CAAAAGGACCTAATAATGGATATTTGGATAAATTTGGTAATGAATGGACT 1530
         ||   :::   ::::::            |||:::      :::||||||
     800 roProSerThrSerSer............LysLeuArgProArgTrpThr 812

    1531 ...AAAGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATGGGATGTTCA 1577
              :::|||      |||   :::|||              |||
     813 PheThrSerProProValThrThrAlaGlnAla.........ThrValPr 826

    1578 ATTGTCTAAAACAGGAAGAGAGCAA.....................:..C 1603
          :::      |||::::::   :::                      |
     826 oValProProThrSerGlnProArgPheSerAsnLeuSerMetLeuValL 843

    1604 TTGGATGGGCTAGT 1617
         ||   |||||||||
     843 euGlnTrpAlaSer 847
```

```
seq_name: SwissProt_40:YAV1_SCHPO

seq_documentation_block:
ID   YAV1_SCHPO     STANDARD;      PRT;  1794 AA.
AC   Q10172;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   01-OCT-1996 (Rel. 34, Last annotation update)
DE   Hypothetical 193.3 kDa protein C27F1.01C in chromosome I.
GN   SPAC27F1.01C OR SPAC25G10.09C.
OS   Schizosaccharomyces pombe (Fission yeast).
OC   Eukaryota; Fungi; Ascomycota; Schizosaccharomycetes;
OC   Schizosaccharomycetales; Schizosaccharomycetaceae;
OC   Schizosaccharomyces.
OX   NCBI_TaxID=4896;
RN   [1]
RP   SEQUENCE OF 1-1748 FROM N.A.
RC   STRAIN=972;
RA   Harris D., McDonald S., Barrell B.G., Rajandream M.A., Walsh S.V.;
RL   Submitted (FEB-1996) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1457-1794 FROM N.A.
RC   STRAIN=972;
RA   McLean J., Harris D., Barrell B.G., Rajandream M.A., Walsh S.V.;
RL   Submitted (APR-1996) to the EMBL/GenBank/DDBJ databases.
CC   -!- SIMILARITY: SOME, TO YEAST PAN1 AND TO MAMMALIAN EPS15.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z69368; CAA93290.1; -.
DR   EMBL; Z70691; CAA94638.1; -.
DR   InterPro; IPR002048; EF-hand.
DR   InterPro; IPR000261; EPS15_repeat.
DR   InterPro; IPR003124; WH2.
DR   Pfam; PF00036; efhand; 3.
DR   Pfam; PF02205; WH2; 1.
DR   SMART; SM00027; EH; 2.
DR   SMART; SM00246; WH2; 1.
KW   Hypothetical protein; Repeat.
SQ   SEQUENCE   1794 AA;  193279 MW;  1F042BA5E80E9E27 CRC64;


alignment_scores:
            Quality:  131.50              Length:    520
              Ratio:  0.632                Gaps:     25
 Percent Similarity:  40.000   Percent Identity: 22.500

alignment_block:
 US-09-303-518D-465 x YAV1_SCHPO   ..
```

```
Align seg 1/1  to: YAV1_SCHPO  from: 1  to: 1794

     114 CCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGG 163
         |||||||:::      |||      :::      ||||||   |||||||::
    1326 ProSerThrThrSerThrSerPheAsnThrAlaProIleProGlnGlnAl 1342

     164 GGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGC 213
         :                                    ::::::   :::|||
    1342 aProLeu.....................GluAsnGlnPheSerLysM 1351

     214 CATCAGTTGGGCAACCTGTTCATCCAGCAGG.............CGGCCA 250
           |||:::      ||||||:::||||||                :::|||
    1351 etSerLeuGluProProValArgProAlaValProThrSerProLysPro 1367

     251 TTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAA 300
                    |||:::::::::   |||   |||
    1368 GlnIleProAspSerSerAsnValHisAlaProPro............Pr 1380

     301 GTCCATTCCCCCTTCGACAACCATGCCT...CACATTCCGATTCTGATGA 347
             :::   |||   ::::::||||||   |||      :::      :
    1380 oProValGlnProMetAsnAlaMetProSerHisAsnAlaValAsnAlaA 1397

     348 AGCCGGTAGTCCCGTTGACGGATTCAGCCTTT.................. 379
         ::|||   :::|||         ::::::|||
    1397 rgProSerAlaProGluArgArgAspSerPheGlySerValSerSerGly 1413

     380 ......ACCGCATCCATTGGGACGGATACGAACACCATCCCG........ 415
               ::::::||||||      :::|||:::|||:::|||
    1414 SerAsnValSerSerIleGluAspGluThrSerThrMetProLeuLysAl 1430

     416 ..............CCGACGGCTATGACGGGCCACAGGGCGGCGGCTA 449
                       |||   |||   ::::::|||         ::::
    1430 aSerGlnProThrAsnProGlyAlaProSerAsnHisAlaProGlnValV 1447

     450 TCCCG..........................CTCCCAAAGGCGCGA 469
         ::|||                                |||||||||
    1447 alProProAlaProMetHisAlaValAlaProValGlnProLysAlaPro 1463

     470 GGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTC 519
         |||:::   |||   :::                        :::   ||
    1464 GlyMetValThrAsnAlaPro.....................AlaProSe 1473

     520 AACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTT.... 565
         |:::   |||:::      ||||||   :::   |||      :::
    1473 rSerAlaProAlaProProAlaProValSerGlnLeuProProAlaValP 1490

     566 ..............TCCACAATACCGGTAGTATGCTGACGCAAGGAG 598
                       |||   ||||||   |||:::   :::
    1490 roAsnValProValProSerMetIleProSerValAlaGlnGln...... 1504

     599 TAGGCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGA 648
                   |||::::::   |||   ||||||||||||   |||
    1505 ......ProProSerSerValAlaProAlaThrAlaProSerSerThr.. 1518

     649 TCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAA 698
               :::|||||||:::   |||:::      ::::::   |||   :
    1519 .......LeuProProSerGlnSerSerPheAlaHisValProSerProA 1533

     699 CATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGG 748
         ::
    1533 la.............................................. 1533

     749 GTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTT 798

    1533 ..............................................  1533

     799 TCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCA 848
          ||||||
    1534 .ProPro.....................................  1535

     849 ACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACC 898
                       |||||||:::|||||||||   :::   |||::::::
```

```
  1536 ..........AlaProGlnHisProSerAlaAlaAlaLeuSerSerAla 1548

   899 CCA............ATGCCGCACAAGGCA..................... 916
       |||            |||||||||::::::
  1549 ProAlaAspAsnSerMetProHisArgSerSerProTyrAlaProGlnGl 1565

   917 .......TAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAA 959
               ||||||:::|||||||:::            :::   :::
  1565 uProValGlnLysProGlnAlaIleAsnAsnIleAlaProAlaThrAsnL 1582

   960 AGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGG......CGGCATCA 1003
         |||    |||   |||   :::   |||:::|||      ::::::
  1582 euGlyThr.SerGlnSerPheSerProArgMetGlyProValAsnAsnSe 1598

  1004 CGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAG......ATCGCATTG 1047
        |||::: |||:::      :::      |||:::      :::|||:::
  1598 rGlySer.ProLeuAlaMetAsnAlaAlaGlyGlnProSerLeuAlaVal 1614

  1048 CCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGC 1097
       |||         ||||||   |||::::::|||         |||
  1615 ProAlaValProSerAlaProSerAsnHisPheAsnProPheAlaLysMe 1631

  1098 CAAATACCCGTCCCCT.........TACCATTCCCGAAATATCCGTTCAA 1138
         :::   |||:::|||               |||   :::      :::|
  1631 tGlnProProAlaProSerProLeuGlnProSerGlyHisAspSerAspA 1648

  1139 ACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCCG 1188
       ||   :::|||:::      :::|||:::      |||:::
  1648 snTrpSerGlnHisGlyAspGluGluGluGluAspSerGluAspAspIle 1664

  1189 CCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGAC 1238
          |||:::         :::   ||||||   |||
  1665 ArgSerSerLysAspAlaAlaAlaLeuAlaAlaLysLeu........... 1677

  1239 CAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAA 1288
               |||   |||
  1678 ..........PheGlyGly............................... 1680

  1289 AATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGAT 1338
                                 :::      ::::::|||:::::
  1681 .........................MetAlaProAlaHisProValSer 1688

  1339 GAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTC 1388
          |||      :::|||:::::::::::            |||
  1689 ThrProProValArgProGlnSerAlaAlaProProGlnMetSer..... 1703

  1389 TATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCA 1438
                              |||      |||            :::
  1704 .....................AlaProThrProProProProProMetS 1713

  1439 GATATATCCCACCTAAAAATTACTCTCCTTCAGCA...CCGCTACCAAAA 1485
               ||||||          ||||||||   |||:::|||
  1713 erValProProPro............ProSerAlaProProMetProAla 1725

  1486 GGACCT 1491
       ||||||
  1726 GlyPro 1727
```

```
seq_name: SwissProt_40:DRPL_HUMAN

seq_documentation_block:
ID   DRPL_HUMAN     STANDARD;      PRT;  1185 AA.
AC   P54259;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Atrophin-1 (Dentatorubral-pallidoluysian atrophy protein).
GN   DRPLA.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Cerebellum, and Brain;
RX   MEDLINE=95144175; PubMed=7842016;
RA   Nagafuchi S., Yanagisawa H., Ohsaki E., Shirayama T., Tadokoro K.,
RA   Inoue T., Yamada M.;
RT   "Structure and expression of the gene responsible for the triplet
RT   repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA).";
RL   Nat. Genet. 8:177-182(1994).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=96262314; PubMed=8965642;
RA   Margolis R.L., Li S.-H., Young W.S., Wagster M.V., Stine O.C.,
RA   Kidwai A.S., Ashworth R.G., Ross C.A.;
RT   "DRPLA gene (atrophin-1) sequence and mRNA expression in human
RT   brain.";
RL   Brain Res. Mol. Brain Res. 36:219-226(1996).
RN   [3]
RP   SEQUENCE OF 470-725 FROM N.A.
RC   TISSUE=Brain cortex;
RX   MEDLINE=93315145; PubMed=8325628;
RA   Li S.-H., McInnis M.G., Margolis R.L., Antonarakis S.E., Ross C.A.;
RT   "Novel triplet repeat containing genes in human brain: cloning,
RT   expression, and length polymorphisms.";
RL   Genomics 16:572-579(1993).
CC   -!- TISSUE SPECIFICITY: THE LEVELS ARE RELATIVELY HIGH IN THE BRAIN,
CC       OVARY, TESTIS AND PROSTATE. LOWER LEVELS ARE DETECTED IN THE
CC       LIVER, THYMUS AND LEUKOCYTES.
CC   -!- POLYMORPHISM: THE POLY-GLN REGION OF DRPLA IS HIGHLY POLYMORPHIC
CC       (7 TO 23 REPEATS) IN THE NORMAL POPULATION AND IS EXPANDED TO
CC       ABOUT 49-75 REPEATS IN DRPLA PATIENTS. LONGER EXPANSIONS RESULT IN
CC       EARLIER ONSET AND MORE SEVERE CLINICAL MANIFESTATIONS OF THE
CC       DISEASE.
CC   -!- DISEASE: DEFECTS IN DRPLA ARE THE CAUSE OF DENTATORUBRAL-
CC       PALLIDOLUYSIAN ATROPHY, AN AUTOSOMAL DOMINANT NEURODEGENERATIVE
CC       DISORDER CHARACTERIZED BY A LOSS OF NEURONS IN THE DENTATE
CC       NUCLEUS, RUBRUM, GLOGUS PALLIDUS AND LUYS'BODY. CLINICAL FEATURES
CC       ARE MYOCLONUS EPILEPSY, DEMENTIA, AND CEREBELLAR ATAXIA. ONSET OF
CC       THE DISEASE OCCURS USUALLY IN THE SECOND DECADE OF LIFE AND DEATH
CC       IN THE FOURTH.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D31840; BAA06626.1; -.
DR   EMBL; U23851; AAB50276.1; -.
DR   EMBL; L10377; -; NOT_ANNOTATED_CDS.
DR   HSSP; P00651; 1LRA.
DR   MIM; 125370; -.
DR   InterPro; IPR002951; Atrophin.
DR   PRINTS; PR01222; ATROPHIN.
KW   Triplet repeat expansion; Polymorphism.
FT   DOMAIN       73     82       SER/GLU-RICH (MIXED CHARGE).
FT   DOMAIN      302    305       POLY-PRO.
FT   DOMAIN      376    382       POLY-SER.
FT   DOMAIN      386    397       POLY-SER.
FT   DOMAIN      442    447       POLY-PRO.
FT   DOMAIN      479    483       POLY-HIS.
FT   DOMAIN      484    497       POLY-GLN.
FT   DOMAIN      504    507       POLY-PRO.
FT   DOMAIN      564    574       POLY-SER.
FT   DOMAIN      704    707       POLY-PRO.
FT   DOMAIN      802    815       ARG/ALA-RICH (MIXED CHARGE).
FT   DOMAIN      816    827       ARG/GLU-RICH (MIXED CHARGE).
FT   DOMAIN      925    934       ARG/GLU-RICH (MIXED CHARGE).
FT   CONFLICT     94     94       MISSING (IN REF. 2).
FT   CONFLICT    333    333       Y -> H (IN REF. 2).
FT   CONFLICT    339    339       M -> I (IN REF. 2).
FT   CONFLICT    541    541       P -> T (IN REF. 3).
```

```
FT   CONFLICT   1028   1028       G -> A (IN REF. 2).
SQ   SEQUENCE   1185 AA;  124785 MW;  56C306267331C005 CRC64;


alignment_scores:
            Quality:  130.00              Length:    685
              Ratio:  0.494                Gaps:     37
 Percent Similarity:  38.394   Percent Identity: 22.336

alignment_block:
 US-09-303-518D-465 x DRPL_HUMAN   ..

 Align seg 1/1  to: DRPL_HUMAN  from: 1  to: 1185

      65 CACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGAC 114
         |||   |||            |||      :::|||         |||
     253 HisProPro.........ProThrThrProIleSerValSerSerSerGl 266

     115 CGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGG 164
          :::|||   :::   ||||||          |||      ::::::||||
     266 yAlaSerGlyAlaProProThrLysProProThrThrProValGlyGlyG 283

     165 GGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGCC 214
         ||||||||||||||||||         :::::::::    |||:::   ||
     283 lyAsnLeuProSerAlaProProProAlaAsnPhe.ProHisValThrPr 299

     215 ATCAGTTGGGCAACCTGTTCATCCAGCAGG.CGGCCATTAAAGGAAATAT 263
         |::::::      |||      ||||||   |||||||||::::::   |
     299 oAsnLeuProProPro......ProAlaLeuArgProLeuAsnAsnAlaS 314

     264 CGGCTACA.........TTGTCCGCTTTTCCGATCACGGGCACGAAGTCC 304
         ||||||::         |||:::|||   |||:::   |||
     314 erAlaSerProProGlyLeuGlyAlaGlnProLeuProGly......... 327

     305 ATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGT 354
                              |||:::|||
     328 .......................HisLeuPro................... 330

     355 AGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGA 404
                      |||   :::      :::::::|||         :
     331 ...............SerProTyr......AlaMetGlyGlnGlyMetG 340

     405 ACACCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCG 454
         ::   :::||||||            |||         |||      |||
     340 lyGlyLeuProProGlyProGluLysGlyProThrLeuAlaProSerPro 356

     455 ......CTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGC 498
               ||||||   |||   :::        |||
     357 HisSerLeuProProAlaSerSerSerAlaProAlaProPro...MetAr 372

     499 GTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACA 548
          :::|||      |||:::|||:::      ::::::|||:::   ::::
     372 gPheProTyrSerSerSerSerSerSerSerAlaAlaAlaSerSerSerS 389

     549 ACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAG 598
         :::::   |||:::   |||:::   |||
     389 erSerSerSerSerSerSerAlaSerProPhe..................... 399

     599 TAGGCGACGGATTCAAACGCGCCACCCGATACAGCC...CCGAGCTGGAC 645
                            |||   :::   |||   ||||||:::
     400 ......................ProAlaSerGlnAlaLeuProSerTyrPr 409

     646 AGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAA 695
                       ||||||         |||:::|||:::::|||::::
     409 oHisSerPheProProPro.........ThrSerLeuSerValSerAsnG 423

     696 AAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGC 745
         ::      :::      |||   :::|||   :::||||||:::
     423 lnProProLysTyrThrGlnProSerLeuProSerGlnAlaVal...... 437

     746 AGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTG 795
                                                 |||
     438 ..........................................TrpSerGl 440

     796 CTTTCCACCGAAAACA.............AGATGGCGCGCATCAACGATTT 833
             ||||||                    |||      |||::::::
     440 nGlyProProProProProProTyrGlyArgLeuLeuAlaAsnSer.... 455

     834 GGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATT 883
                       |||:::          |||      ||||||:::
     456 ..............AsnAlaHisProGlyProPheProProSerThrGly 467

     884 GGGCAGTCCAAAACCCCAATGCCG.............CACAAGGCATAGAA 921
         :::||||||   :::      |||            ||| :::|||::::
     468 AlaGlnSerThrAlaHisProProValSerThrHisHis.HisHisHisG 484

     922 GCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGC 971
         :::::||||||            ::::::   :::|||:::      :::
     484 lnGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnHisHisGly 500

     972 TGTTCGGGGAAAATACGGCTTGGG........................... 995
             ||||||            |||
     501 AsnSerGlyProProProProGlyAlaPheProHisProLeuGluGlyGl 517

     996 ....CGGCATCACGGCACATCCT........................... 1014
                 ||||||  |||||||||
     517 ySerSerHisHis.AlaHisProTyrAlaMetSerProSerLeuGlySer 533

    1014 ................................................... 1014

     534 LeuArgProTyrProProGlyProAlaHisLeuProProProHisSerGl 550

    1015 .GTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCG 1063
          |||:::   ||||||   |||      :::   |||   ::::::|||:
     550 nValSerTyrSerGlnAlaGlyProAsnGlyProProValSerSerSerS 567

    1064 CC...GTCAGCGACAATTTTGCCGATGCGGCATAC.............. 1095
         ::      |||:::::::   :::::::::::|||
     567 erAsnSerSerSerSerThrSerGlnGlySerTyrProCysSerHisPro 583

    1096 .......................GCCAAATACCCGTCCCCTTACCATTC 1121
                                |||   ||||||   |||    :::|
     584 SerProSerGlnGlyProGlnGlyAlaProTyrProPhePro.ProValP 600

    1122 CCGAAATA......TCCGTTCAAACTTGGAGCAGCGTTACGGCA..... 1159
         ||   :::      |||      |||  |||:::|||   |||
     600 roThrValThrThrSerSerAlaThrLeuSerThrValIleAlaThrVal 616

    1160 .....................AAGAAAACATCACCTCCTCAACCGTGCCG 1188
                              |||   :::|||||||||   |||  |||
     617 AlaSerSerProAlaGlyTyrLysThrAlaSerProProGlyPro..Pro 632

    1189 CCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCG..... 1233
         |||                          :::||||||   |||
     633 Pro.......................TyrGlyLysArgAlaProSerPr 641

    1234 ..........AAGACCAAAGTGCCGTTTGACGGTAAAGGGTTT....... 1266
                  ||||||   :::|||            |||:::
     641 oGlyAlaTyrLysThrAlaThrPro.........ProGlyTyrLysProG 655

    1267 ........CCGAATTTTGAAAAAGACGTAAAATACGATACGAGA...... 1302
                 |||:::|||:::      :::             |||
     655 lySerProProSerPheArgThrGlyThrProProGlyTyrArgGlyThr 671

    1303 .......................ATTAATACCGCTGTACCACAAGTGAA 1328
                                 ::::::   :::   |||   |||::
     672 SerProProAlaGlyProGlyThrPheLysProGlySerProThrValGl 688

    1329 TCCTATAGATGAACCCGTCTTTAATCCTAAAGGT.............. 1362
         :|||         |||      :::|||:::|||
     688 yProGlyProLeuProProAlaGlyProSerGlyLeuProSerLeuProP 705

    1363 ..........TCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGA 1401
                    :::   :::|||:::        |||   |||
     705 roProProAlaAlaProAlaSerGlyProProLeuSerAlaThr...... 719
```

```
    1402 ATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACC 1451
                                      :::|||:::      |||
     720 .............................GlnIleLysGlnGluProAl 726

    1452 TAAAAATTAC...TCTCCTTCAGCACCGCTACCA......AAAGGACCTA 1492
          ::::::|||   :::|||:::::::|||:::|||      ::::::|||:
     726 aGluGluTyrGluThrProGluSerProValProProAlaArgSerProS 743

    1493 AT.............................................. 1494
          ::
     743 erProProProLysValValAspValProSerHisAlaSerGlnSerAla 759

    1495 ......AATGGATATTTGGATAAATTTGGTAAT................. 1521
                |||   :::|||||||:::      |||
     760 ArgPheAsnLysHisLeuAspArgGlyPheAsnSerCysAlaArgSerAs 776

    1521 .............................................. 1521

     776 pLeuTyrPheValProLeuGluGlySerLysLeuAlaLysLysArgAlaA 793

    1521 .............................................. 1521

     793 spLeuValGluLysValArgArgGluAlaGluGlnArgAlaArgGluGlu 809

    1522 ....................GAATGGACTAAAGGTCCATCAAGAACTAA 1550
                             |||      |||      :::|||   ||
     810 LysGluArgGluArgGluArgGluArgGluLysGluArgGluArgGluLy 826

    1551 AGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACAGGAAGA 1596
          |   :::|||:::|||   :::|||:::|||::::::   ||||||
     826 sGluArgGluLeuGluArgSerValLysLeuAlaGlnGluGlyArg 841
```

seq_name: SwissProt_40:VGLX_HSVEB

seq_documentation_block:

```
ID   VGLX_HSVEB     STANDARD;      PRT;   797 AA.
AC   P28968;
DT   01-DEC-1992 (Rel. 24, Created)
DT   01-DEC-1992 (Rel. 24, Last sequence update)
DT   01-DEC-1992 (Rel. 24, Last annotation update)
DE   Glycoprotein X precursor.
GN   71.
OS   Equine herpesvirus type 1 (strain Ab4p) (EHV-1).
OC   Viruses; dsDNA viruses, no RNA stage; Herpesviridae;
OC   Alphaherpesvirinae; Varicellovirus.
OX   NCBI_TaxID=31520;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92295566; PubMed=1318606;
RA   Telford E.A.R., Watson M.S., McBride K., Davison A.J.;
RT   "The DNA sequence of equine herpesvirus-1.";
RL   Virology 189:304-316(1992).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M86664; AAB02506.1; -.
DR   PIR; H36802; VGBEX1.
KW   Glycoprotein; Transmembrane; Signal.
FT   SIGNAL        1     22       POTENTIAL.
FT   CHAIN        23    797       GLYCOPROTEIN X.
FT   DOMAIN       23    465       SER/THR-RICH.
FT   TRANSMEM    766    790       POTENTIAL.
FT   CARBOHYD    590    590       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   797 AA;  80342 MW;  50C9ED9211F5E5B2 CRC64;
```

alignment_scores:

```
           Quality:  128.50              Length:    565
             Ratio:  0.547                Gaps:     18
 Percent Similarity:  41.593   Percent Identity: 20.177
```

alignment_block:

US-09-303-518D-465 x VGLX_HSVEB ..

Align seg 1/1 to: VGLX_HSVEB from: 1 to: 797

```
    110 TCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGC 159
        ||||||:::::   ::::::   |||:::::::::::   |||:::::
     22 SerThrThrThrThrGluThrThrThrSerSerSerSerThrSerGlySe 38

    160 AGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACA 209
        :|||                                           :
     38 rGly.........................................G 40

    210 AAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAA 259
        :::::   |||   :::|||   :::||||||      |||      :::
     40 lnSerThrSerSerGlyThrThrAsnSerSerSerSerProThrThrSer 56

    260 ATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCC 309
            :::|||   |||:::   |||   :::          |||   ||
     57 ProProThrThrSerSerSerProProThrSerThrHisThrSerSerPr 73

    310 CCCTTCGACA...ACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAG 356
        |   ||||||   |||:::                         :::      :
     73 oSerSerThrSerThrGlnSerSerSerThrAlaAlaThrSerSerSerA 90

    357 TCCCGTTGACGGATTCAGCCTTTACCGCA.........TCCATTGGGACG 397
        ::|||   |||   |||:::   |||:::         |||   :::|||
     90 laProSerThrAlaSerSerThrThrSerIleProThrSerThrSerThr 106

    398 GATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGC 447
        :::|||:::|||      |||||||||   |||         ::::::::
    107 GluThrThrThrThrThrProThrAlaSerThrThrThrProThrThrTh 123

    448 TATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGG 497
        :          |||   :::   :::      ||||||:::|||      |
    123 rThrAlaAlaProThrThrAlaAlaThrThrThrAlaValThrThrAlaA 140

    498 CG.....................TTGCCCAAAATATCCGCCTCAACCTGA 526
        ||                                 :::|||::::::
    140 laSerThrSerAlaGluThrThrThrAlaThrAlaThrAlaThrSerThr 156

    527 CCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACC 576
        |||||||||:::::|||   :::      :::|||::::::|||
    157 ProThrThrThrThrProThrSerThrThrThrThrThrAlaThrThrTh 173

    577 GGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCG 626
         |||                         |||   :::::::::      |
    173 rVal.....................ProThrThrAlaSerThrThrThrA 183

    627 ATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACG 676
        |||||:::   :::   |||      |||                ::::|||
    183 spThrThrThrAlaAlaThrThrThrAlaAlaThrThrThrAlaAlaThr 199

    677 GCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTC 726
        :::      ::::::   ::::::::::::          ::
    200 ThrThrAlaAlaThrThrThrAlaAlaThrThrThrAlaAlaThrThrTh 216

    727 GGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGT 776
        :|||   :::      :::   :::   :::::::|||   |||
    216 rAlaAlaThrThrThrAlaAlaThrThrSerSerAlaThrThrAlaAlaT 233

    777 TATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCA 826
             ||||||   :::               |||         |||:::
    233 hrThrThrAlaAlaThrThrThrAlaAlaThrThr.....ThrAlaAla 247

    827 ACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCA.. 874
        |||                  ::::::   |||         :::|||
    248 ThrThrThrAlaAlaThrThrThrAlaAlaThrThrThrGlySerProTh 264
```

```
  874 ..................................................  874

  264 rSerGlySerThrSerThrThrGlyAlaSerThrSerThrProSerAlaS  281

  875 .....................................TCCGCGATTGGG  886
                                                ||||||   :::
  281 erThrAlaThrSerAlaThrProThrSerThrSerThrSerAlaAlaAla  297

  887 CAGTCCAAAACCCCAATGCCGCACAAGGCA.....................  916
      ::::::||||||   |||   :::|||
  298 ThrThrSerThrProThrProThrSerAlaAlaThrSerAlaGluSerTh  314

  917 .....................................TAGAAGCCGTCAG  929
                                                  ||||||
  314 rThrGluAlaProThrSerThrProThrThrAspThrThrThrProSerG  331

  930 CAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGG  979
          :::         :::|||||||
  331 luAlaThrThrAlaThrThrSerPro.......................  339

  980 GAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGCAG 1029
      |||:::|||:::   :::||||||             ::::::
  340 GluSerThrThrValSerAlaSerThrThrSerAlaThrThrThrAlaPh  356

 1030 ATGGGCGAGATCGCATTGCCGAAAGGGAAAT..................... 1060
          ::::::||||||          :::
  356 eThrThrGluSerHisThrSerProAspSerSerThrGlySerThrSerT  373

 1061 .....CCGCCGTCAGCG......ACAATTTTGCCGATGCGGCATACGCCA 1099
               ||||||::      |||:::   |||         ||||||
  373 hrAlaGluProSerSerThrPheThrLeuThrProSerThrAlaThrPro  389

 1100 AATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAG 1149
      :::|||      :::|||      :::   |||:::::::::   |||::
  390 SerThrAspGlnPheThrGlySerSerAlaSerThrGluSerAspSerTh  406

 1150 CGT..................TACGGCAAAGAAAACATCACCTCCTCAA 1180
      :                     |||   |||:::||||||:::|||:
  406 rAspSerSerThrValProThrThrGlyThrGluSerIleThrGluSerS  423

 1181 CCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCAC 1230
      ::      :::   :::   |||      |||::::::::::   :::
  423 erSerThrThrGluAlaSerThrAsn......LeuGlySerSerThrTyr  437

 1231 CCGAAGACCAAA..................................... 1242
         :::|||:::
  438 GluSerThrGluAlaLeuGluThrProAspGlyAsnThrThrSerGlyAs  454

 1243 ....GTGCCGTTTGACGGTAAAGGGTTTCCGAATTTT...GAAAAAGACG 1285
          :::|||      :::         |||:::|||   :::   :::
  454 nThrThrProSerProSerProArgThrProSerPheAlaAspThrGlnG  471

 1286 TAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATA 1335
              |||:::   ::::::|||             :::|||
  471 lnThrProAspAsnGlyValSerThrGlnHisThrThrIleAsn......  485

 1336 GATGAACCCGTCTTTAATCCTAAA..........................  1359
      |||:::   :::   |||   :::
  486 AspHisThrThrAlaAsnAlaGlnLysHisAlaGlyHisHisArgGlyAr  502

 1360 .......GGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAA 1402
             |||      ||||||   ::::::   |||   |||:::
  502 gAlaGlyGlyArgArgGlySerProGlnGlyGlySerHisThrThr....  517

 1403 TTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCT 1452
                     |||:::      |||:::         |||
  518 ................ProHisProAspArgLeuThrProSerProAsp  528

 1453 AAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAAT 1497
          :::|||:::   ::::::      |||:::|||   ||||||
  529 AspThrTyrAspAspAspThrAsnHisProAsnGlyArgAsnAsn 543

seq_name: SwissProt_40:FIG2_YEAST
```

```
seq_documentation_block:
ID   FIG2_YEAST     STANDARD;      PRT;  1609 AA.
AC   P25653;
DT   01-MAY-1992 (Rel. 22, Created)
DT   01-MAY-1992 (Rel. 22, Last sequence update)
DT   15-DEC-1998 (Rel. 37, Last annotation update)
DE   Factor induced gene 2.
GN   FIG2 OR YCR089W OR YCR89W OR YCR1102.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92397594; PubMed=1523889;
RA   Wilson C., Grisanti P., Frontali L.;
RT   "The complete sequence of a 6146 bp fragment of Saccharomyces
RT   cerevisiae chromosome III contains two new open reading frames.";
RL   Yeast 8:569-575(1992).
CC   -!- FUNCTION: REQUIRED FOR EFFICIENT MATING.
CC   -!- INDUCTION: BY MATING PHEROMONES.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X59720; CAA42254.1; -.
DR   PIR; S19504; S19504.
DR   PIR; S25345; S25345.
DR   SGD; S0000685; FIG2.
SQ   SEQUENCE   1609 AA;  166049 MW;  7D66AD7F85A7B852 CRC64;


alignment_scores:
            Quality:  131.50              Length:    566
              Ratio:  0.460                Gaps:     25
 Percent Similarity:  50.530   Percent Identity: 20.848

alignment_block:
 US-09-303-518D-465 x FIG2_YEAST   ..

 Align seg 1/1  to: FIG2_YEAST  from: 1  to: 1609

      17 AAATATCCCTTATTCTGTCCATACTGGCAGTGTGCCTGCCGATGCATGCA 66
         ::::|||      :::||||||           |||           ::
     763 GluTyrThrThrTrpCysPro............AlaSerSerIleAlaTy 775

      67 CACGCCTCAGATT...TGGCAAACGATTCTTTTATCCGGCAGGTTCTCGA 113
         :|||   :::|||   :::::::|||:::::::|||:::   :::     :
     775 rThrThrSerIleSerTyrLysThrLeuValLeuThrThrGluValCysS 792

     114 CCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGG 163
         ::   |||      :::|||||||      |||:::   :::|||::::::
     792 erHisSerGluCysThrProThrValIleThrSerValThrAlaThrSer 808

     164 GGGAACTTGCCG..........................AGCGCAGCGGT 186
         :::::::::|||                          ::::::::
     809 SerThrIleProLeuLeuSerThrSerSerSerThrValLeuSerSerTh 825

     187 CATATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCAT 236
         ::::|||:::               |||   |||   :::|||   ::::
     825 rValSerGluGlyAlaLysAsnProAlaAlaSerGluValThrIleAsnT 842

     237 CCAGCAGGCGGCCATTAAAGGAAATATCGGCTACA.........TTGTCC 277
         :::::           :::|||   ::::::|||         :::|||
     842 hrGlnValSerAlaThrSerGluAlaThrSerThrSerThrGlnValSer 858

     278 GCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGCC 327
```

```
     |||         |||:::|||   |||          |||:::|||
 859 AlaThrSerAlaThrAlaThrAlaSer......GluSerSerThrThrSe 873

 328 TCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCT 377
      ::::::          :::   :::      |||      :::   |
 873 rGlnValSerThrAlaSerGluThrIleSerThrLeuGlyThrGlnAsnP 890

 378 TTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATG 427
     |||||             |||      :::   :::   |||:::
 890 heThr............ThrThrGlySerLeuLeuPheProAlaLeuSer 902

 428 ACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATA 477
     |||             :::::::::          |||:::      |||
 903 ThrGluMetIleAsnThrThrValValSerArgLysThrLeuIleIleSe 919

 478 TACA......GCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAA 521
      |||      ::::::      |||::::::|||   :::   :::::::
 919 rThrGluValCysSerHisSerLysCysValProThrValIleThrGluV 936

 522 CCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCG............ 559
     ::      :::   ::::::|||:::||||||   |||
 936 alValThrSerLysGlyThrProSerAsnGlyHisSerSerGlnThrLeu 952

 560 .....ACCGTTTCCACAATACCGGTAGTATGCTG 588
                  |||:::|||:::      :::|||
 953 GlnThrGluAlaValGluValThrLeuSerSerHisGlnThrValThrMe 969

 589 ACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGCCCCGA 638
           |||   :::::::::|||   :::  |||  :::      |
 969 tSerThrGluValCysSerAsnSerIleCysThrProThrValIleThrS 986

 639 GCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATA 688
     ||         |||:::   |||          |||:::
 986 erValGlnMetArgSerThrProPheProTyrLeuThrSerSerThrSer 1002

 689 TCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGAT 738
     |||||||:::   :::|||:::::::::::::::|||:::|||::::::
1003 SerSerSerLeuAlaSerThrLysLysSerSerLeuGluAlaSerSerGl 1019

 739 GCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTT 788
                  :::::::::|||:::|||   |||      |||:
1019 uMetSerThrPheSerValSerThrGlnSerLeuProLeu......AlaP 1034

 789 GGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAG 838
     :::::|||                ::::::   |||::::::|||     :::
1034 heThrCys..........SerGluLysArgSerThrThrSerValSer 1046

 839 ATATGGCGCAAC............TCAAAGACTATGCCGCAGCAGCCATC 876
      |||   |||                ::::      |||   ::::::   ::
1047 GlnTrpSerAsnThrValLeuThrAsnThrIleMetSerSerSerSerAs 1063

 877 CGCGATTGGGCAGTCCAAAACCCCA..................... 901
     ::::|||:::   ::::::   |||
1063 nValIleSerThrAsnGluLysProSerSerThrThrSerProTyrAsnP 1080

 902 .................ATGCCGCACAAGGCATAGAAGCCGTCAGCAATA 934
                      :::|||   ::::::      ||||||
1080 heSerSerGlyTyrSerLeuProSerSerSer...ThrProSerGlnTyr 1095

 935 TCTTTACGGCAGTCATCCCCGTCA...AAGGGATTGGAGCTGTTCGGGGA 981
     ||||||      ::::::   :::   :::|||::::::
1096 SerLeuSerThrAlaThrThrThrIleAsnGlyIleLysThrVal..... 1110

 982 AAATACGGCTTGGGCG......GCATCACGGCACATCCTGTCAAGCGGTC 1025
      |||:::|||:::      |||::::::      ::::::::::::
1111 .TyrThrThrTrpCysProLeuAlaGluLysSerThrValAlaAlaSerS 1127

1026 GCAGATGGGCGAGATCGCATTGCCGA......AAGGGAAATCCGCCGTCA 1069
       :::   :::|||||||      |||      :::::::::   ||||||
1127 erGlnSerSerArgSerValAspArgPheValSerSerSerLysProSer 1143

1070 GCGACAATT......TTGCCGATGCGGCATACGCCAAATACCCGTCCCCT 1113
     :::::::::           :::::::::|||   :::|||
```

```
1144 SerSerLeuSerGlnThrSerIleGlnTyrThrLeuSerThrAlaThrTh 1160

1114 TACCATTCCCGAAATATCCGTT......CAAACTTGGAGCAGCGTTACGG 1157
      ||||||      :::::::::          ||||||      :::|||:
1160 rThrIleSerGlyLeuLysThrValTyrThrThrTrpCysProLeuThrS 1177

1158 CAAAGAAAACATC..............ACCTCCTCAACC......... 1182
     ::|||:::|||                |||||||||||
1177 erLysSerThrLeuGlyAlaThrThrGlnThrSerSerThrAlaLysVal 1193

1183 ............GTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAA 1220
                  :::      :::::::::   :::::::::|||:::::
1194 ArgIleThrSerAlaSerSerAlaThrSerThrSerIleSerLeuSerTh 1210

1221 CAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGA 1270
     ::::   :::   :::::::::      :::   :::||||||
1210 rSerThrGluSerGluSerSerSerGlyTyrLeuSerLysGly....... 1224

1271 ATTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCA 1320
                |||        |||:::   :::      ||||||
1225 ..............ValCysSerGlyThrGluCysThrGlnAspValPro 1236

1321 ......CAAGTGAATCCTATAGATGAACCCGTCTTTAATCCTAAAGGTTC 1364
           :::   :::|||   :::      :::::::::|||:::   ||
1237 ThrGlnSerSerSerProAlaSerThrLeuAlaTyrSerProSerValSe 1253

1365 TGTCGGATCGGCTCATTCTTGGTCTATAACT................... 1395
     |::::::|||:::   |||:::|||   |||
1253 rThrSerSerSerSerPheSerThrThrThrAlaSerThrLeuThrS 1270

1396 ..GCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATAT 1443
       ::::::   :::         ||||||   ::::::   |||
1270 erThrHisThrSerValProLeuLeuProSerSerSerSerIleSerAla 1286

1444 ATCCCACCTAAAAAT.........TACTCTCCTTCAGCACCGCTACCA 1482
           |||::::::            |||   |||   |||   |||
1287 SerSerProSerSerThrSerLeuLeuSerThrSerLeuProSerPro 1302
```

```
seq_name: SwissProt_40:ENAM_PIG

seq_documentation_block:
ID   ENAM_PIG       STANDARD;      PRT;  1142 AA.
AC   O97939;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Enamelin precursor.
GN   ENAM.
OS   Sus scrofa (Pig).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Cetartiodactyla; Suina; Suidae; Sus.
OX   NCBI_TaxID=9823;
RN   [1]
RP   SEQUENCE FROM N.A., AND CHARACTERIZATION.
RC   TISSUE=Enamel epithelium;
RX   MEDLINE=98040070; PubMed=9372788;
RA   Hu C.-C., Fukae M., Uchida T., Qian Q., Zhang C.H., Ryu O.H.,
RA   Tanabe T., Yamakoshi Y., Murakami C., Dohi N., Shimizu M.,
RA   Simmer J.P.;
RT   "Cloning and characterization of porcine enamelin mRNAs.";
RL   J. Dent. Res. 76:1720-1729(1997).
RN   [2]
RP   SEQUENCE OF 39-773 FROM N.A., SEQUENCE OF 39-49; 174-276;
RP   515-524; 535-578; 641-646; 663-665; 670-686; 740-750; 765-773 AND
RP   833-848, AND CHARACTERIZATION.
RX   MEDLINE=97350624; PubMed=9206327;
RA   Fukae M., Tanabe T., Murakami D., Dohi N., Uchida T., Shimizu M.;
RT   "Primary structure of porcine 89 kda enamelin.";
RL   Adv. Dent. Res. 10:111-118(1996).
CC   -!- FUNCTION: PEPTIDES DERIVED FROM THE PARENT ENAMELIN ARE COMPONENTS
CC       OF ENAMEL, A UNIQUE AND HIGHLY MINERALIZED ECTODERMAL TISSUE
CC       COVERING VERTEBRATE TEETH.
CC   -!- TISSUE SPECIFICITY: EXPRESSED BY SECRETORY-PHASE AMELOBLASTS.
```

```
CC       INTACT ENAMELIN AND LARGE-MOLECULAR-WEIGHT ENAMELINS ARE LIMITED
CC       TO THE MOST SUPERFICIAL LAYER OF THE DEVELOPING ENAMEL MATRIX,
CC       WHILE LOW-MOLECULAR-WEIGHT ENAMELINS ARE OBSERVED IN DEEPER
CC       ENAMELIN. PREFERENTIAL LOCALIZATION AMONG THE CHRYSTALLITES IN ROD
CC       AND INTERROD ENAMEL.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED FROM LATE DIFFERENTIATION TO THE
CC       TRANSITION STAGE.
CC   -!- PTM: PARENT ENAMELIN IS PROTEOLYTICALLY CLEAVED INTO SEVERAL
CC       SMALLER POLYPEPTIDES. CLEAVAGE OF N-TERMINAL REGION OF ENAMELIN
CC       OCCURS SOON AFTER SECRETION.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U52196; AAD10837.1; -.
DR   GlycoSuiteDB; O97939; -.
KW   Signal; Enamel; Glycoprotein; Hydroxylation; Phosphorylation.
FT   SIGNAL        1     38
FT   CHAIN        39   1142       ENAMELIN.
FT   CHAIN        39      ?       56 KDA ENAMELIN.
FT   CHAIN        39    665       89 KDA ENAMELIN.
FT   CHAIN        39      ?       142 KDA ENAMELIN.
FT   CHAIN        39      ?       155 KDA ENAMELIN.
FT   CHAIN       174    276       32 KDA ENAMELIN.
FT   CHAIN       515    665       25 KDA ENAMELIN.
FT   CHAIN       670      ?       34 KDA ENAMELIN.
FT   CHAIN         ?      ?       45 KDA ENAMELIN.
FT   MOD_RES      53     53       PHOSPHORYLATION (PROBABLE).
FT   MOD_RES     191    191       PHOSPHORYLATION.
FT   MOD_RES     216    216       PHOSPHORYLATION.
FT   MOD_RES     547    547       HYDROXYLATION.
FT   CARBOHYD    245    245       N-LINKED (GLCNAC. . .).
FT   CARBOHYD    252    252       N-LINKED (GLCNAC. . .).
FT   CARBOHYD    264    264       N-LINKED (GLCNAC. . .).
FT   CARBOHYD    291    291       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    462    462       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    929    929       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1040   1040       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT    680    680       H -> D (IN REF. 2).
FT   CONFLICT    838    840       RDH -> TTI (IN REF. 2).
SQ   SEQUENCE   1142 AA;  128352 MW;  938306BC87CC5FC6 CRC64;


alignment_scores:
            Quality:  127.00            Length:    609
              Ratio:  0.477              Gaps:     40
 Percent Similarity:  43.678   Percent Identity: 23.317

alignment_block:
 US-09-303-518D-465 x ENAM_PIG   ..

 Align seg 1/1  to: ENAM_PIG  from: 1  to: 1142

      3 GGGCATTTCCCG...CAAAATATCCCTTATTCTGTCCATACTGGCAGTGT 49
        |||||||||       :::|||   |||:::   :::|||   |||:::::
     60 GlyHisPheAsnPheMetAsnAlaProHisMetAlaHisLeuGlyThrLe 76

     50 GCCTGCCGATGCATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATC 99
        :   :::::::::            :::|||      ::::::      |
     76 uTyrGlyAsnGlyMetGlnLeuProGlnPhePheProGlnTyrGlnMetP 93

    100 CGGCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCT 149
        ||          :::|||         :::   :::   :::   |||
     93 roMetTrpProGlnProProProAsnLysLysHisProGlnLysProSer 109

    150 ATT...CGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGAT 196
              :::||||||:::   |||      |||   :::|||
    110 AlaSerLysGlnGlnSerLysThrAspProAlaProGluSer..... 123

    197 TGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGC....A 242
                |||||||:::       ||||||      |||       :
    124 .........GlnLysProAsnGlnProGlnProLysThrProThrProL 137

    243 GGCGGCCATTAAAGGAAATATCGGCTACATTGTCC.............. 277
        :::::||||||:::|||   |||   |||   :::
    137 ysGlnProLeuAsnGluProSerProThrProThrGlnProGluGluGlu 153

    278 ..............GCTTTTCCGATCACGGGCACGAAGTCCATTCCCCC 312
                      |||||||||       ||| ::::::   ||||||
    154 ThrGlnThrProGlnAlaPheProProPheGly.AsnGlyLeuPheProT 170

    313 TTCGACAACCATGCCTCACAT........................... 333
          :::||||||       |||
    170 yrGlnGlnProLeuTrp.HisValProHisArgIleProProGlyTyrGl 186

    334 ..........TCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCA 373
                 :::::::::::::|||:::|||:::|||       ||||||
    186 yArgProProThrSerAsnGluGluGlyGlyAsnProTyrPheGlyPheP 203

    374 GCCTTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGC 423
                                                        |||
    203 he.........................................Gly 204

    424 TATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGA 473
        |||   |||         ||||||   |||   |||      ::::::::
    205 TyrHisGly......PheGlyGlyArgPro...ProTyrTyrSerGluGl 218

    474 TATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACC 523
        ::::::::::   |||::: :::         |||   :::   :::
    218 uMetPheGluGlnAspPhe.Glu.........LysProLysGluLysAsp 231

    524 TGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAAT 573
          |||   |||   :::|||   :::       |||:::   :::|||::
    232 ProProLysThrGluThrProAlaThrGluProSerValAsnThrThrVa 248

    574 ACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCA. 622
        :|||                :::       :::|||:::   |||||||||
    248 lPro..............GluThrAsnSerThrGlnProAsnAlaProA 260

    623 ...........CCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCG 661
                    ||||||:::|||:::   |||:::         |||
    260 snProArgGlyAsnAspThrSerProThrGlyThrSerGlyGlnGlyPro 276

    662 ...CCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGC 708
          |||:::   :::   :::   |||                 :::::
    277 AsnProArgSerAsnProThrGlyGln.................AsnGl 287

    709 GCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGA 758
        :         ::::::|||:::|||::::::|||              :
    287 yProAlaValAsnValSerGlyGlnGlyValPro..............A 299

    759 AGGCTCAAACATTGCTGTTATGCACGGCTTGG................. 790
        :::::|||              :::   |||
    299 rgSerGln..............SerProTrpGlyProArgGlnThrIle 310

    791 ............................GTCTGCTTTCCACCGAAAAC 810
                                                |||||   :::
    311 IleHisGluAsnTyrProAsnProAsnIleArgGlyPheProAlaArgAr 327

    811 AAGATGGCGCG.....................CATCAACGATTTGGCAGA 839
          :::|||                       |||:::|||   |||
    327 gGlnTrp.ArgProProGlyProAlaMetGlyHisArgArgAsnGlyPro 343

    840 TATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAG 889
        :::      :::|||::::::   |||:::      |||||| |||
    344 PheTyrArgAsnGlnGlnIleGlnArgGly......ProArg.Trp.... 356

    890 TCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATCTTT 939
                                                  |||   |||
    357 ....................................AsnSerPhe 359
```

```
 940 ACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATAC.. 987
     |||            ::::::|||      |||||||||      |||
 360 Thr............LeuGluGlyLysGlnAlaValArgProGlyTyrPr 372

 988 ..........GGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGC 1027
                ::   |||   ||||||:::            :::
 372 oThrTyrArgArgValTyrGlySerThrAlaArgSerAsnProProAsnT 389

1028 AGATGGGCGAGATCGCA............TTGCCGAAAGGG......AAA 1059
       |||:::  |||              |||:::|||     |||
 389 yrAlaGlyAsnSerAlaAsnLeuArgArgLysProGluGlyProAsnLys 405

1060 TCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCAAATACCCGTC 1109
     :::  :::     |||  |||                  |||::
 406 AsnProMetValThrAsnValAlaPro.................ProGl 416

1110 CCCTTACCAT..............TCCCGAAATATCCGTTCAAACTTGG 1144
     :|||  |||              ::::::||||||:::::      |
 416 yProLysHisGlyThrValAspGlnAsnGluAsnIleGlnAsnProArgG 433

1145 AGCAGCGTTACGGC...AAAGAAAACATCACCTCCTCAACCGTGCCGCCG 1191
     ||:::::   :::  ||||||:::  :::     |||      |||
 433 luLysGlnValSerGlnLysGluArgThrValValProThrArgAspPro 449

1192 TCA............AACGGAAAGAATGTGAAACTGGCAAACAAACGCCA 1229
     |||           |||:::::::::      :::  ||||||  ::
 450 SerGlyProTrpArgAsnSerGlnAspTyrGlyIle...AsnLysSerAs 465

1230 CCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAA 1279
     :  |||         |||  |||:::        |||||||||:::
 465 nTyrLysLeuProGlnProGluAspAsnMetLeuValProAsnPheAsn. 481

1280 AAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAAT 1329
           :::  |||  |||  |||:::     |||:::  :::
 482 ........SerIleAspGlnArgGluAsnSerTyrTyrProArgGlyGlu 495

1330 CCTATAGATGAACCCGTCTTTAAT.....................CC 1355
                |||      :::                      ||
 496 SerLysArgAlaProAsnSerAspGlyGlnThrGlnThrGlnIleIlePr 512

1356 TAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTC 1405
     |||||||  |||  ::  ::                  ||||||
 512 oLysGlyIleValLeuGluProArg.................ArgIleP 523

1406 AATACGCAAAATTACCAAGGCAAGGTAGAATCAGA.........TATATC 1446
       |||  :::     :::|||  :::::::::        |||
 523 roTyrGluSerGluThrAsnGlnProGluLeuLysHisSerAlaTyrGln 539

1447 CCACCTAAAAATTACTCTCCTTCAGCACCG.................CT 1478
     |||        |||:::  :::  |||                   ::
 540 ProVal......TyrThrGluGlyIleProSerProAlaLysGluHisPh 554

1479 ACCAAAAAGGACCTAATAATGGATATTTGGATAAATTTGGTAATGAATGGA 1528
     :|||  |||  |||:::        ::::::::::::     :::
 554 eProAlaGlyArgAsnThrTrpAsnGlnGlnGluIleSerProProPheL 571

1529 CTAAAGGTCCATCAAGA 1545
       :::  |||:::|||
 571 ysGluAspProGlyArg 576
```

```
seq_name: SwissProt_40:KN8R_YEAST

seq_documentation_block:
ID   KN8R_YEAST     STANDARD;      PRT;   893 AA.
AC   P53739;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Probable serine/threonine-protein kinase YNR047W (EC 2.7.1.-).
GN   YNR047W OR N3449.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RA   Pohl T.M.;
RL   Submitted (MAY-1996) to the EMBL/GenBank/DDBJ databases.
CC   -!- SIMILARITY: BELONGS TO THE SER/THR FAMILY OF PROTEIN KINASES.
CC       KIN82 SUBFAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z71662; CAA96328.1; -.
DR   HSSP; Q00534; 1BI8.
DR   SGD; S0005330; YNR047W.
DR   InterPro; IPR000719; Euk_pkinase.
DR   InterPro; IPR002290; Ser_thr_pkinase.
DR   Pfam; PF00069; pkinase; 1.
DR   SMART; SM00220; S_TKc; 1.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; FALSE_NEG.
DR   PROSITE; PS00108; PROTEIN_KINASE_ST; 1.
DR   PROSITE; PS50011; PROTEIN_KINASE_DOM; 1.
KW   Hypothetical protein; Transferase; Serine/threonine-protein kinase;
KW   ATP-binding.
FT   DOMAIN      496    777       PROTEIN KINASE.
FT   NP_BIND     502    510       ATP (BY SIMILARITY).
FT   BINDING     525    525       ATP (BY SIMILARITY).
FT   ACT_SITE    621    621       BY SIMILARITY.
SQ   SEQUENCE   893 AA;  100546 MW;  26AF74EE956F80DB CRC64;


alignment_scores:
            Quality:  126.50              Length:    493
              Ratio:  0.541                 Gaps:     26
 Percent Similarity:  47.465   Percent Identity: 20.892

alignment_block:
 US-09-303-518D-465 x KN8R_YEAST   ..

 Align seg 1/1  to: KN8R_YEAST  from: 1  to: 893

  98 TCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCAC 147
     |||:::|||||||||   ::::::           :::||||||:::::
  84 SerSerArgPheSerLysLeuLysSerMetPheGlnSerGlyAsnSerSe 100

 148 CTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATT 197
     :   :::|||::::::   |||      ::::::::::::   :::|||
 100 rLysAsnAlaSerAlaHisAsnSerSerGlnSerSerLeuGluGlyAsp. 116

 198 GGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGG 247
                          |||:::  ||||||:::  |||
 117 ......................SerAlaSerSerSerSerLysLeuArg 125

 248 CCATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCAC 297
       ::||||  ::::::::::::  :::|||  |||
 126 TyrValLysProMetThrSerValAlaAsnAlaSerPro........... 138

 298 GAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATGA 347
        |||  |||  |||  |||:::|||:::      :::|||:::
 139 .AlaSerProProLeuSerProThrIleProGluThrAspValLeuGlnT 155

 348 AGCCG..................................................... 352
       |||
 155 hrProLysMetValHisIleAspGlnHisGluHisGluArgGluHisSer 171

 353 ............GTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCAT 390
                    :::  |||:::::|||:::|||:::
 172 AsnCysGlySerProIleMetLeuSerSerSerSerPheSer........ 185
```

```
391 TGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGG 440
                          |||   :::|||   |||||| ||||||
186 .....................ProThrValAlaArgThrGly.ThrGly 194

441 CGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACA 490
    |||
195 Arg.............................................. 195

491 TAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGC 540
        ||||||   |||:::   |||   :::|||:::::::::::   ::::
196 ....ArgArgSerProSerThrProIleMetProSerGlnAsnSerAsnA 211

541 AC......CGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTAT 584
    ::         |||:::|||   ||||||      :::|||||||   :::
211 snSerSerSerThrSerAlaIleArgPro...AsnAsnTyrArgHisHis 226

585 GCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGCC 634
    :::   :::::::::         :::::::::::::::|||:::::::
227 SerGlySerGlnGly.........PheSerSerAsnAsnProPheArgGl 240

635 CCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCA 684
     |||||||||   :::           ||||||   :::   :::
240 uArgAlaGlyThrVal............ArgSerSerAsnProTyrPheA 253

685 GATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGG 734
      |||           ::: |||          :::   :::::::::
253 laTyr...........Gln.GlyLeuProThrHisAlaMetSerSerHi 265

735 CGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACG 784
     |||      :::|||:::::::      :::|||   :::         |
265 sAspLeuAspGluGlyPheGlnProTyrAlaAsnGlySer.......G 279

785 GCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTTG 834
    ||:::   :::|||||||||   :::      ::::::   |||:::|||
279 lyIleHisPheLeuSerThrProThr......SerLysThrAsnSerLeu 293

835 GCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTG 884
    ::::::     :::|||::::::  :::          |||::::::
294 ThrAsnThrLysAsnLeuSerAsnLeuSerLeuAsnGluIleLysGlu.. 309

885 GGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAAT. 933
                              :::   ::::::      :::|||
310 .....................AsnGluGluValGlnGluPheAsnAsnG 319

934 .....ATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGG 978
         :::|||         ||||||   |||   ::::::
319 luAspPhePhePheHisAspIlePro...LysAspLeuSer......... 331

979 GGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGCA 1028
                                              :::|||   :::
332 .................................LeuLysAspThrLe 336

1029 GATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATT 1078
        |||:::   :::      ::::::||||||   :::      :::
336 uAsnGlySerProSerArgGlySerSerLysSerProThrIleThrGlnT 353

1079 TTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCCCGAAAT 1128
           :::      :::   :::   :::   |||::::::   |||
353 hrPheProSerIleIleValGlyPheAspAsnGluTyrGluGluAspAsn 369

1129 ATCCGT............TCAAACTTGGAGCAGCGTTACGGCAAAGAAAA 1166
       :::           ::::::   |||||||:::         :::||
370 AsnAsnAspLysHisAspGluLysGluGluGlnGlnThrThrThrAspAs 386

1167 CATCACCTCCTCAACCGTGCCG...CCGTCAAACGGAAAGAATGTGAAAC 1213
     |   |||   :::      |||      :::|||||||||
386 nLysThrArgAsnLeuSerProThrLysGlnAsnGlyLys.......... 399

1214 TGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGG 1263
                  ||||||:::   |||:::|||
400 ........AlaThrHisProArgIleLysIlePro.............. 408

1264 TTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACCGC 1313
```

```
                                              ::::::   ||
409 .....................................LeuArgArgAl 412

1314 TGTACCACAAGTGAATCCTATAGAT...GAACCCGTCTTTAATCCTAAAG 1360
     |:::   :::   |||   ::::::          :::   :::|||   :
412 aAlaSerGluProAsnGlyLeuGlnLeuAlaSerAlaThrSerProThrS 429

1361 GTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTCAATAC 1410
     ::|||      ||||||::::::   |||   :::::::::|||:::
429 erSer......SerAlaArgLysThrSerGlySerSerAsnIleAsn... 442

1411 GCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCTAAAAATTA 1460
        |||:::|||   |||:::               ||||||:::::::
443 AspLysIleProGlyGlnSer.........ValProProProAsnSerPh 456

1461 CTCTCCTTCAGCACCGCTACCAAAA 1485
     :   |||:::   |||   ||||||
456 ePheProGlnGluProSerProLys 464
```

```
seq_name: SwissProt_40:GTFB_STRMU

seq_documentation_block:
ID   GTFB_STRMU     STANDARD;      PRT;  1476 AA.
AC   P08987; O69381; O69384; O69387; O69390; O69396;
DT   01-NOV-1988 (Rel. 09, Created)
DT   15-JUL-1999 (Rel. 38, Last sequence update)
DT   15-JUL-1999 (Rel. 38, Last annotation update)
DE   Glucosyltransferase-I precursor (EC 2.4.1.5) (GTF-I) (Dextransucrase)
DE   (Sucrose 6-glucosyltransferase).
GN   GTFB.
OS   Streptococcus mutans.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Streptococcaceae;
OC   Streptococcus.
OX   NCBI_TaxID=1309;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=GS-5;
RX   MEDLINE=87308013; PubMed=3040685;
RA   Shiroza T., Ueda S., Kuramitsu H.K.;
RT   "Sequence analysis of the gtfB gene from Streptococcus mutans.";
RL   J. Bacteriol. 169:4263-4270(1987).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=MT4239, MT4245, MT4251, MT4467, AND MT8148;
RX   MEDLINE=98231643; PubMed=9570124;
RA   Fujiwara T., Terao Y., Hoshino T., Kawabata S., Ooshima T., Sobue S.,
RA   Kimura S., Hamada S.;
RT   "Molecular analyses of glucosyltransferase genes among strains of
RT   Streptococcus mutans.";
RL   FEMS Microbiol. Lett. 161:331-336(1998).
CC   -!- FUNCTION: PRODUCTION OF EXTRACELLULAR GLUCANS, THAT ARE THOUGHT
CC       TO PLAY A KEY ROLE IN THE DEVELOPMENT OF THE DENTAL PLAQUE BECAUSE
CC       OF THEIR ABILITY TO ADHERE TO SMOOTH SURFACES AND MEDIATE THE
CC       AGGREGATION OF BACTERIAL CELLS AND FOOD DEBRIS.
CC   -!- CATALYTIC ACTIVITY: Sucrose + {(1,6)-alpha-D-glucosyl}(N) = D-
CC       fructose + {(1,6)-alpha-D-glucosyl}(N+1).
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- DISEASE: DENTAL CARIES.
CC   -!- MISCELLANEOUS: GTF-I SYNTHESIZES WATER-INSOLUBLE GLUCANS (ALPHA
CC       1,3-LINKED GLUCOSE AND SOME 1,6 LINKAGES), GTF-S SYNTHESIZES
CC       WATER-SOLUBLE GLUCANS (ALPHA 1,6-GLUCOSE). GTF-SI SYNTHESIZES BOTH
CC       FORMS OF GLUCANS.
CC   -!- SIMILARITY: TO OTHER GLUCOSYLTRANSFERASES AND SOME TO A GLUCAN-
CC       BINDING PROTEIN FROM S.MUTANS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
```

```
DR   EMBL; M17361; AAA88588.1; -.
DR   EMBL; D88651; BAA26101.1; -.
DR   EMBL; D88654; BAA26105.1; -.
DR   EMBL; D88657; BAA26109.1; -.
DR   EMBL; D88660; BAA26113.1; -.
DR   EMBL; D89977; BAA26119.1; -.
DR   PIR; B33135; B33135.
DR   InterPro; IPR002479; CW_binding.
DR   InterPro; IPR003318; Glyco_hydro_70.
DR   Pfam; PF01473; CW_binding_1; 13.
DR   Pfam; PF02324; Glyco_hydro_70; 1.
KW   Transferase; Glycosyltransferase; Signal; Repeat; Dental caries.
FT   SIGNAL       1     34       POTENTIAL.
FT   CHAIN       35   1476       GLUCOSYLTRANSFERASE-I.
FT   DOMAIN      35   1051       CATALYTIC (APPROXIMATE).
FT   DOMAIN    1097   1476       GLUCAN-BINDING (APPROXIMATE).
FT   REPEAT    1097   1130       A REPEAT.
FT   DOMAIN    1161   1470       5 X TANDEM REPEATS.
FT   REPEAT    1161   1210       1.
FT   REPEAT    1225   1275       2.
FT   REPEAT    1290   1340       3.
FT   REPEAT    1355   1405       4.
FT   REPEAT    1420   1470       5.
FT   VARIANT     62     62       S -> T (IN STRAIN MT4239).
FT   VARIANT     65     65       I -> T (IN STRAINS MT4239, MT4245,
FT                               MT4251, MT4467 AND MT8148).
FT   VARIANT     68     68       A -> V (IN STRAIN MT4239).
FT   VARIANT     78     78       Q -> P (IN STRAIN MT4251).
FT   VARIANT     86     86       S -> I (IN STRAIN MT4239).
FT   VARIANT     89     89       S -> F (IN STRAIN MT4251).
FT   VARIANT    168    168       K -> N (IN STRAIN MT4251).
FT   VARIANT    276    276       D -> S (IN STRAINS MT4239, MT4245 AND
FT                               MT4251).
FT   VARIANT    399    399       N -> R (IN STRAIN MT4239).
FT   VARIANT    474    474       I -> T (IN STRAIN MT4239).
FT   VARIANT    512    512       K -> R (IN STRAIN MT8148).
FT   VARIANT    519    519       F -> Y (IN STRAIN MT8148).
FT   VARIANT    701    701       T -> I (IN STRAIN MT8148).
FT   VARIANT    708    708       A -> V (IN STRAIN MT8148).
FT   VARIANT    938    938       F -> L (IN STRAIN MT8148).
FT   VARIANT    952    957       YGTPVA -> FGKPVE (IN STRAINS MT4245,
FT                               MT4251 AND MT8148).
FT   VARIANT    963    964       NT -> SV (IN STRAINS MT4245, MT4251 AND
FT                               MT8148).
FT   VARIANT    968    970       VDG -> ADS (IN STRAINS MT4245, MT4251 AND
FT                               MT8148).
FT   VARIANT   1086   1086       A -> T (IN STRAIN MT4239).
FT   VARIANT   1158   1158       S -> N (IN STRAIN MT4239).
FT   VARIANT   1163   1163       H -> Y (IN STRAIN MT4251).
FT   VARIANT   1168   1168       E -> K (IN STRAIN MT8148).
FT   VARIANT   1182   1182       Y -> C (IN STRAIN MT8148).
FT   VARIANT   1234   1234       A -> P (IN STRAIN MT4239).
FT   VARIANT   1263   1263       H -> P (IN STRAIN MT8148).
FT   VARIANT   1263   1263       H -> R (IN STRAINS MT4239, MT4245 AND
FT                               MT4251).
FT   VARIANT   1264   1264       H -> Y (IN STRAINS MT4245 AND MT4251).
FT   VARIANT   1272   1272       G -> S (IN STRAINS MT4245 AND MT4251).
FT   VARIANT   1329   1329       Y -> H (IN STRAINS MT4239, MT4245, MT4251
FT                               AND MT8148).
FT   VARIANT   1394   1394       H -> Y (IN STRAINS MT4245 AND MT4251).
FT   VARIANT   1402   1402       G -> S (IN STRAINS MT4245 AND MT4251).
FT   VARIANT   1459   1459       Y -> H (IN STRAIN MT4467).
FT   CONFLICT   570    570       R -> A (IN REF. 1).
FT   CONFLICT   800    817       ADQDVRVAASTAPSTDGK -> LIKMFALRLARPHQQMA
FT                               (IN REF. 1).
FT   CONFLICT  1310   1310       H -> L (IN REF. 1).
SQ   SEQUENCE  1476 AA;  165685 MW;  3479B62B07694D98 CRC64;
```

```
alignment_scores:
            Quality:  125.50              Length:    610
              Ratio:  0.468                Gaps:     33
 Percent Similarity:  43.934   Percent Identity: 21.639
```

```
alignment_block:
 US-09-303-518D-465 x GTFB_STRMU   ..

 Align seg 1/1  to: GTFB_STRMU  from: 1  to: 1476

     64 GCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGA 113
        ::::::|||      |||:::|||   |||   :::   :::|||
    617 SerTyrAlaLeuLeuLeuThrAsnLysSerSerValProArgValTyrTy 633

    114 CCGTCAGCATTTCGAACCCGACGGGAAATAC..................... 144
           :::   |||      ||||||:::|||
    633 rGlyAspMetPheThrAspAspGlyGlnTyrMetAlaHisLysThrIleA 650

    145 .....................................CACCTATTC 153
                                             ::::::
    650 snTyrGluAlaIleGluThrLeuLeuLysAlaArgIleLysTyrValSer 666

    154 GGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCAT.............. 189
        |||:::::::::         ::::::   |||:::
    667 GlyGlyGlnAlaMetArgAsnGlnGlnValGlyAsnSerGluIleIleTh 683

    190 ..........ATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACC 229
                     |||   |||   :::::::::   :::   |||:::
    683 rSerValArgTyrGlyLysGlyAlaLeuLysAlaThrAspThrGlyAspA 700

    230 TGTTCATCCAGCAG.........GCGGCCATTAAAGGAAATATCGGCTAC 270
              :::            |||:::|||:::||||||
    700 rgThrThrArgThrSerGlyValAlaValIleGluGlyAsnAsnProSer 716

    271 ATT...GTCCGCTTTTCCGAT.........................CA 290
        :::   ::::::   ||||||                         ||
    717 LeuArgLeuLysAlaSerAspArgValValValAsnMetGlyAlaAlaHi 733

    291 CGGGCACGAAGTCCATTCCCCCTTC............GACAAC......C 322
        |   ::::::::::::   |||:::          ||||||
    733 sLysAsnGlnAlaTyrArgProLeuLeuLeuThrThrAspAsnGlyIleL 750

    323 ATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTC 372
          |||   ||||||||||:::::   ||||||
    750 ysAlaTyrHisSerAspGlnGluAlaAlaGly.................. 760

    373 AGCCTTTACCGCATCCATTGGGACGGATACGAA......CACCATCCCGC 416
           |||   |||          |||      |||                ||
    761 ...LeuValArgTyrThrAsnAspArgGlyGluLeuIlePheThrAlaAl 776

    417 CGAC......GGCTATGACGGGCCACAGGGCGGCGGCTATCCC....... 453
        ||||      ||||||   :::||||||   :::||||||
    776 aAspIleLysGlyTyrAlaAsnProGlnValSerGlyTyrLeuGlyValT 793

    454 ..GCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGTT 501
          :::|||   ||||||                              ::::::
    793 rpValProValGlyAla.........................AlaAla 800

    502 GCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAACG 551
           |||::::::|||:::      :::          ||||||      ::
    801 AspGlnAspValArgValAlaAlaSerThrAlaProSerThrAspGlyLy 817

    552 GCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAG 601
        :   |||          |||:::::::::::::::   ::::::   :::
    817 sSerVal.........HisGlnAsnAlaAlaLeuAspSerArgValMetP 831

    602 GCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCG 651
          :::||||||::::::         ::::::   :::   ::::::
    831 heGluGlyPheSerAsnPheGlnAlaPheAlaThrLysLysGluGluTyr 847

    652 GGCAATGCCGCCGAAGCTTTCAAC.....................GGCAC 680
           |||::::::   |||   |||                      |||::
    848 ThrAsnValValIleAlaLysAsnValAspLysPheAlaGluTrpGlyVa 864

    681 TGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCG 730
        ::::|||               :::      |||   :::   |||::::
    864 lThrAsp..............PheGluMetAlaProGlnTyrValSerS 876
```

```
731 CAGGCGATGCCGTG.........CAGGGTATAAGCGAAGGCTCAAACATT 771
    ::  |||:::           :::   |||:::::|||      :::
876 erThrAspGlySerPheLeuAspSerValIleGlnAsnGlyTyrAlaPhe 892

772 GCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCG 821
    :::     :::   ||||||   :::|||     ||||||   :::
893 ThrAspArgTyrAspLeuGly...IleSerLysProAsnLysTyrGlyTh 908

822 CATCAACGATTTGGCAGAT...ATGGCGCAACTCAAAGACTATGCCGCAG 868
      :::|||||:::      :::      |||   :::   :::
908 rAlaAspAspLeuValLysAlaIleLysAlaLeuHisSerLysGlyIleL 925

869 CAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATA 918
    ::::::   ||||||:::   ::::::      |||
925 ysValMetAlaAspTrpValProAspGlnMetTyrAlaProGluLys 941

919 GAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGG 968
    |||:::|||:::
942 GluValValThr..................................... 945

969 AGCTGTTCGG...GGAAAATACGGCTTGGGCGGCATCACGGCACATCCTG 1015
     |||:::|||      |||||||||                    ||||
946 .AlaThrArgValAspLysTyrGlyThr.................ProV 956

1016 TCAAGCGGTCGCAGATG...GGCGAGATCGCATTGCCGAAAGGGAAATCC 1062
     ||      ||||||:::   :::   :::   :::      |||||||||
956 alAlaGlySerGlnIleLysAsnThrLeuTyrValValAspGlyLysSer 972

1063 GCCGTCAGCGACAAT.....................TTTGCCGATGCGGC 1091
     :::   :::|||:::                    |||   :::
973 SerGlyLysAspGlnGlnAlaLysTyrGlyGlyAlaPheLeuGluGluLe 989

1092 ATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACT 1141
       ||||||||||||:::   :::      ::::::|||
989 uGlnAlaLysTyrProGluLeuPheAlaArgLysGlnIle..... 1002

1142 TGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCCG... 1188
                                    |||:::   ||||||
1003 ..............................SerThrGlyValProMet 1008

1189 ...CCGTCA................................AACGGAAA 1202
        ||||||                              ||||||
1009 AspProSerValLysIleLysGlnTrpSerAlaLysTyrPheAsnGlyTh 1025

1203 GAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTG 1252
     |||:::
1025 rAsnIleLeu..................................... 1028

1253 ACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGA 1302
       |||:::|||              ||||||        :::|||
1029 ..GlyArgGlyAlaGlyTyrValLeuLysAspGlnAlaThrAsnThrTyr 1044

1303 ATTAATACCGCT...GTACCACAAGTGAATCCTATAGATGAACCCGTCTT 1349
     :::|||   :::        ::::::|||   :::   :::   :::::
1045 PheAsnIleSerAspAsnLysGluIleAsnPheLeuProLysThrLeuLe 1061

1350 TAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCA 1399
     :|||      ::::::||||||   ::::::::: |||
1061 uAsnGlnAspSerGlnValGlyPheSerTyrAsp.GlyLysGlyTyrVal 1077

1400 GAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCA 1449
       ::::::|||::::::||||||::|||   ::::::   ||||||::
1078 TyrTyrSerThrSerGlyTyrGlnAlaLys...AsnThrPheIleSerGl 1093

1450 CCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAATGG 1499
     :                                              |
1093 u.............................................GlyA 1095

1500 ATATTTGGATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAGAACTA 1549
     ||   |||      ::      ||||||      |||
1095 spLysTrpTyrTyrPheAspAsnAsnGlyTyrMetValThrGlyAlaGln 1111

1550 AAGGTCAAGAATTTGAATGGGATGTTC......AATTGTCTAAAACAGGA 1593
     ::::::::::::::|||      |||      |||   |||:::   ::
1112 SerIleAsnGlyValAsnTyrTyrPheLeuSerAsnGlyLeuGlnLeuAr 1128

1594 AGAGAGC.........AACTTGGATGGG 1612
     :::::::         |||   ||||||
1128 gAspAlaIleLeuLysAsnGluAspGly 1137
```

seq_name: SwissProt_40:SON_HUMAN

seq_documentation_block:

ID SON_HUMAN STANDARD; PRT; 2426 AA.
AC P18583; O95981; Q9UPY0; Q14120; O14487; Q9UKP9; Q9H7B1; Q9P070;
AC Q9P072;
DT 01-NOV-1990 (Rel. 16, Created)
DT 01-MAR-2002 (Rel. 41, Last sequence update)
DT 01-MAR-2002 (Rel. 41, Last annotation update)
DE SON protein (SON3) (Negative regulatory element-binding protein) (NRE-
DE binding protein) (DBP-5) (Bax antagonist selected in saccharomyces 1)
DE (BASS1) (Protein C21orf50).
GN SON OR NREBP OR DBP5 OR C21ORF50 OR KIAA1019.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A. (ISOFORMS A; B; C; D; E AND F).
RX MEDLINE=21564202; PubMed=11707072;
RA Reymond A., Friedli M., Neergaard Henrichsen C., Chapot F.,
RA Deutsch S., Ucla C., Rossier C., Lyle R., Guipponi M.,
RA Antonarakis S.E.;
RT "From PREDs and open reading frames to cDNA Isolation: revisiting the
RT Human Chromosome 21 Transcription Map.";
RL Genomics 78:46-54(2001).
RN [2]
RP SEQUENCE FROM N.A. (ISOFORM G).
RC TISSUE=Liver;
RX MEDLINE=21316479; PubMed=11306577;
RA Sun C.-T., Lo W.-Y., Wang I.-H., Lo Y.-H., Shiou S.-R., Lai C.-K.,
RA Ting L.-P.;
RT "Transcription repression of human hepatitis B virus genes by negative
RT regulatory element-binding protein/SON.";
RL J. Biol. Chem. 276:24059-24067(2001).
RN [3]
RP SEQUENCE OF 1-689 FROM N.A. (ISOFORM H).
RC TISSUE=Placenta;
RA Casadei R., Strippoli P., D'Addabbo P., Canaider S., Lenzi L.,
RA Vitale L., Giannone S., Carinci P., Zannotti M.;
RL Submitted (OCT-2001) to the EMBL/GenBank/DDBJ databases.
RN [4]
RP SEQUENCE OF 1-130 FROM N.A.
RC TISSUE=Smooth muscle;
RA Kawakami T., Noguchi S., Itoh T., Shigeta K., Senba T., Matsumura K.,
RA Nakajima Y., Mizuno T., Morinaga M., Tanigami A., Fujiwara T., Ono T.,
RA Yamada K., Fujii Y., Ozaki K., Hirao M., Ohmori Y., Ota T., Suzuki Y.,
RA Obayashi M., Nishi T., Shibahara T., Tanaka T., Nakamura Y.,
RA Isogai T., Sugano S.;
RT "NEDO human cDNA sequencing project.";
RL Submitted (AUG-2000) to the EMBL/GenBank/DDBJ databases.
RN [5]
RP SEQUENCE OF 1-114 FROM N.A.
RC TISSUE=Blood;
RA Ye M., Zhang Q.H., Zhou J., Shen Y., Wu X.Y., Guan Z.Q., Wang L.,
RA Fan H.Y., Mao Y.F., Dai M., Huang Q.H., Chen S.J., Chen Z.;
RT "Human partial CDS from cd34+ stem cells.";
RL Submitted (MAY-1999) to the EMBL/GenBank/DDBJ databases.
RN [6]
RP SEQUENCE OF 437-2426 FROM N.A. (ISOFORM B).
RC TISSUE=Brain;
RX MEDLINE=99397452; PubMed=10470851;
RA Kikuno R., Nagase T., Ishikawa K.-I., Hirosawa M., Miyajima N.,
RA Tanaka A., Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. XIV.

```
RT   The complete sequences of 100 new cDNA clones from brain which code
RT   for large proteins in vitro.";
RL   DNA Res. 6:197-205(1999).
RN   [7]
RP   SEQUENCE OF 554-2426 FROM N.A. (ISOFORM A).
RX   MEDLINE=92049296; PubMed=1944255;
RA   Chumakov I.M., Berdichevskii F.B., Sokolova N.V., Reznikov M.V.,
RA   Prasolov V.S.;
RT   "Identification of a protein product of a novel human gene SON and
RT   the biological effect upon administering a changed form of this gene
RT   into mammalian cells.";
RL   Mol. Biol. (Mosk) 25:731-740(1991).
RN   [8]
RP   SEQUENCE OF 709-1079 FROM N.A. (ISOFORM I).
RC   TISSUE=Placenta;
RX   MEDLINE=93062885; PubMed=1435774;
RA   Bliskovskii V.V., Kirillov A.V., Zakhariev V.M., Chumakov I.M.;
RT   "The human son gene: the large and small transcripts contains various
RT   5'-terminal sequences.";
RL   Mol. Biol. (Mosk) 26:807-812(1992).
RN   [9]
RP   SEQUENCE OF 1009-1131 FROM N.A.
RC   TISSUE=Placenta;
RX   MEDLINE=93062884; PubMed=1435773;
RA   Bliskovskii V.V., Berdichevskii F.B., Tkachenko A.V., Belova M.E.,
RA   Chumakov I.M.;
RT   "Coding part of the son gene small transcript contains four areas of
RT   complete tandem repeats.";
RL   Mol. Biol. (Mosk) 26:793-806(1992).
RN   [10]
RP   SEQUENCE OF 1145-2426 FROM N.A. (ISOFORM F).
RX   MEDLINE=93048367; PubMed=1424986;
RA   Mattioni T., Hume C.R., Konigorski S., Hayes P., Osterweil Z.,
RA   Lee J.S.;
RT   "A cDNA clone for a novel nuclear protein with DNA binding
RT   activity.";
RL   Chromosoma 101:618-624(1992).
RN   [11]
RP   SEQUENCE OF 1692-2175 FROM N.A. (ISOFORM A).
RX   MEDLINE=89039788; PubMed=3054499;
RA   Berdichevskii F.B., Chumakov I.M., Kiselev L.L.;
RT   "Decoding of the primary structure of the son3 region in human
RT   genome: identification of a new protein with unusual structure and
RT   homology with DNA-binding proteins.";
RL   Mol. Biol. (Mosk) 22:794-801(1988).
RN   [12]
RP   SEQUENCE OF 1939-2426 FROM N.A. (ISOFORM J).
RC   TISSUE=Cerebellum;
RX   MEDLINE=99439804; PubMed=10509013;
RA   Greenhalf W., Lee J., Chaudhuri B.;
RT   "A selection system for human apoptosis inhibitors using yeast.";
RL   Yeast 15:1307-1321(1999).
CC   -!- FUNCTION: Represses hepatitis B virus (HBV) core promoter activity
CC       and transcription of HBV genes and production of HBV virions.
CC       Binds to the consensus DNA sequence: 5'-GA[GT]AN[CG][AG]CC-3'.
CC       Might protect cells from apoptosis. Might be involved in pre-mRNA
CC       splicing (By similarity).
CC   -!- SUBCELLULAR LOCATION: Nuclear with a speckled distribution.
CC   -!- ALTERNATIVE PRODUCTS: 10 isoforms; A, B, C, D, E, F (shown here),
CC       G, H, I and J; may be produced by alternative splicing.
CC   -!- TISSUE SPECIFICITY: Widely expressed, with the higher expression
CC       seen in leukocyte and heart.
CC   -!- DOMAIN: Contains 8 types of repeats which are distributed in 3
CC       regions.
CC   -!- MISCELLANEOUS: Colocalizes with the pre-mRNA splicing factor
CC       SFRS2/SC-35.
CC   -!- SIMILARITY: CONTAINS 1 G-PATCH DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 DRBM (DOUBLE-STRANDED RNA-BINDING) DOMAIN.
CC   -!- CAUTION: ISOFORM A SEQUENCE FROM REF.7 DIFFERS FROM THAT SHOWN
CC       DUE TO A FRAMESHIFT.
CC   -!- CAUTION: ISOFORM F SEQUENCE FROM REF.10 DIFFERS FROM THAT SHOWN
CC       DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF380179; AAL34497.1; -.
DR   EMBL; X63753; CAA45282.1; ALT_FRAME.
DR   EMBL; M36428; AAA36624.1; -.
DR   EMBL; AF380180; AAL34498.1; -.
DR   EMBL; AF380181; AAL34499.1; -.
DR   EMBL; AF380182; AAL34500.1; -.
DR   EMBL; AF380183; AAL34501.1; -.
DR   EMBL; AF380184; AAL34502.1; -.
DR   EMBL; AY026895; AAK07692.1; -.
DR   EMBL; AF435977; AAL30810.1; -.
DR   EMBL; X63751; CAC69885.1; -.
DR   EMBL; AB028942; BAA82971.1; -.
DR   EMBL; X63071; CAA44793.1; ALT_FRAME.
DR   EMBL; AF139897; AAD50078.1; -.
DR   EMBL; S47238; AAB23945.1; -.
DR   EMBL; AK024752; BAB14985.1; -.
DR   EMBL; AF161428; AAF28988.1; -.
DR   EMBL; AF161430; AAF28990.1; -.
DR   PIR; PN0099; PN0099.
DR   MIM; 182465; -.
DR   InterPro; IPR001159; DS_RBD.
DR   InterPro; IPR000467; G_patch.
DR   Pfam; PF00035; dsrm; 1.
DR   Pfam; PF01585; G-patch; 1.
DR   SMART; SM00358; DSRM; 1.
DR   SMART; SM00443; G-patch; 1.
DR   PROSITE; PS50137; DS_RBD; 1.
DR   PROSITE; PS50174; G_PATCH; 1.
KW   RNA-binding; DNA-binding; Nuclear protein; Repeat;
KW   Alternative splicing.
FT   DOMAIN      726    895       17 X 10 AA TANDEM REPEATS OF L-A-[ST]-
FT                                [NSG]-[TS]-MDSQM.
FT   DOMAIN      912    988       11 X 7 AA TANDEM REPEATS OF [DR]-P-Y-R-
FT                                [LI][AG][QHP].
FT   DOMAIN     1006   1126       14 X 6 AA REPEATS OF [ED]-R-S-M-M-S.
FT   REPEAT     1006   1011       1-1.
FT   REPEAT     1014   1019       1-2.
FT   REPEAT     1021   1026       1-3.
FT   REPEAT     1030   1035       1-4.
FT   REPEAT     1038   1043       1-5.
FT   REPEAT     1046   1051       1-6.
FT   REPEAT     1055   1060       1-7.
FT   REPEAT     1063   1068       1-8.
FT   REPEAT     1071   1076       1-9.
FT   REPEAT     1080   1085       1-10.
FT   REPEAT     1089   1094       1-11.
FT   REPEAT     1100   1105       1-12.
FT   REPEAT     1111   1116       1-13.
FT   REPEAT     1121   1126       1-14.
FT   DOMAIN     1147   1179       3 X 11 AA TANDEM REPATS OF P-P-L-P-P-E-E-
FT                                P-P-[TME]-[MTG].
FT   DOMAIN     1359   1390       4 X 8 AA TANDEM REPEATS OF V-L-E-SS-
FT                                [AVT]-VT.
FT   DOMAIN     1925   1994       7 X 7 AA REPEATS OF P-S-R-R-S-R-[TS].
FT   REPEAT     1925   1931       2-1.
FT   REPEAT     1953   1959       2-2.
FT   REPEAT     1960   1966       2-3.

alignment_scores:
            Quality:  125.50              Length:    589
              Ratio:  0.519                 Gaps:     29
 Percent Similarity:  41.087   Percent Identity: 21.902

alignment_block:
 US-09-303-518D-465 x SON_HUMAN   ..
```

Align seg 1/1 to: SON_HUMAN from: 1 to: 2426

```
 102 GCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTAT 151
     |||   :::|||   |||      :::::::::|||:::
1847 AlaArgLysArgSerSerLysSerLysSerHisArgSer..........  1859

 152 TCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGA 201
       |||       :::      |||:::::|||      |||   ::::
1860 ....GlnThrArgSerArgSerArgSerArgArgArgArgArgSerSerA 1875

 202 AACATACAAAGCCATCAGTTGGGCAA..................... 227
     ::      :::   |||   |||:::
1875 rgSerArgSerLysSerArgGlyArgArgSerValSerLysGluLysArg 1891

 228 .........CCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCT 268
             |||   |||   :::   ::::::   ||||||   :::
1892 LysArgSerProLysHisArgSerLysSerArgGluArgLysArgLys.. 1907

 269 ACATTGTCCGCTTTTCCGATCA.........................C 291
                     ||||||                        |
1908 ..............ArgSerSerSerArgAspAsnArgLysThrValA 1919

 292 GGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTC 341
     ||||||||||      |||   |||:::
1919 rgAlaArgSerArgThrProSerArgArg..................... 1928

 342 TGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATT 391

1928 ..................................................  1928

 392 GGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGC 441
      :::|||   :::||||||||||||||||||      |||:::::|||
1929 .SerArgSerHisThrProSerArgArgArgArgSerArgSerValGlyA 1945

 442 GGCGGCTATCCCGCTCC.........CAAAGGCGCGAGGGATATATACAG 482
     |||||   |||   |||         :::|||||||:::
1945 rgArgArgSerPheSerIleSerProSerArgArgSer............ 1957

 483 CTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACA 532
       |||         |||      :::   |||   :::   :::|||
1958 ...ArgThrProSerArgArgSerArgThrProSerArgArgSerArgTh 1973

 533 ACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGT 582
      |||:::::|||:::   :::            |||:::   |||
1973 rProSerArgArgSerArgThr............ProSerArgArgSerA 1986

 583 ATGCTGACGCAAGG..............AGTAGGCGACGGATTCAAACG 617
       :::   :::|||              ||||||||||||   ::::::
1986 rgThrProSerArgArgSerArgThrProSerArgArgArg...ArgSer 2001

 618 CGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAG 667
     |||      ::::::   |||:::   :::|||:::         |||
2002 ArgSerValValArgArgArgSerPheSerIleSerProValArgLeu.. 2017

 668 CTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGA 717
                                            ||||||:::|||
2018 ..................................ArgArgSerArgT 2022

 718 GAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAA 767
             |||||||||      |||:::      :::|||:::
2022 hrProLeuArgArgArgPheSerArgSerProIleArgArgLys...... 2036

 768 CATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGG 817
                   |||      |||:::   :::   ::::::
2037 ..............ArgSerArgSerSerGluArgGlyArgSerPro.L 2048

 818 CGCGCATCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCA 867
       |||::::::||||||   |||   |||||||||   :::   |||
2048 ysArgLeuThrAspLeu...AspLysAlaGlnLeuLeuGluIleAlaLys 2063

 868 GCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCA........ 909
     |||                                 |||||||||
```

```
2064 Ala.............................AsnAlaAlaAlaMetCy 2070

 910 .......CAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCC 952
           |||:::           |||:::     |||       |
2070 sAlaLysAlaGlyValProLeuProProAsnLeuLysProAlaProProP 2087

 953 CCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGC......TTGGGC 996
     ||              :::::::::   |||   :::         |||
2087 ro..............ThrIleGluGluLysValAlaLysLysSerGly 2098

 997 GGCATCACGGCA..................................... 1008
     |||   |||
2099 GlyAlaThrIleGluGluLeuThrGluLysCysLysGlnIleAlaGlnSe 2115

1008 ..................................................  1008

2115 rLysGluAspAspAspValIleValAsnLysProHisValSerAspGluG 2132

1009 .......................CATCCTGTCAAGCGGTCGCAGATG 1032
                            ||||||   |||   |||:::
2132 luGluGluGluProProPheTyrHisHisProPheLysLeuSerGluPro 2148

1033 GGCGAGATCGCATTG.................................CC 1049
           |||   :::                                ||
2149 LysProIlePhePheAsnLeuAsnIleAlaAlaAlaLysProThrProPr 2165

1050 GAAAGGGAAATCCGCCGTCAGCGACAATTTT................... 1080
     ||||::::::   :::::::::   :::|||
2165 oLysSerGlnValThrLeuThrLysGluProProValSerSerGlySerG 2182

1081 .................GCCGATGCGGCATACGCCAAATACCCGTCCCCT 1113
                     ||||||:::::::|||:::::::::
2182 lnHisArgLysLysGluAlaAspSerValTyrGlyGluTrpValPro... 2197

1114 TACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGA 1163
                  ::::::::::|||   |||:::::          ::
2198 ..............ValGluLysAsnGlyGluGluAsnLysAspAspAs 2209

1164 AAACATCACCTCCTCAACCGTGCCGCCG..............TCAAACG 1198
     :|||:::   ||||||:::::|||                  |||:::
2209 pAsnValPheSerSerAsnLeuProSerGluProValAspIleSerThrA 2226

1199 GAAAGAATGTGAAA...CTGGCAAACAAACGCCACCCGAAGACCAAAGTG 1245
     ::   :::   :::   ||||||:::|||||||      ::::::
2226 laMetSerGluArgAlaLeuAlaGlnLysArgLeuSerGluAsnAlaPhe 2242

1246 CCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGA 1295
        :::::::::   ::::::   |||               :::   ::
2243 AspLeuGluAlaMetSerMetLeuAsn..............ArgAlaGl 2254

1296 TACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCG 1345
     :   ||||||:::   |||      |||:::|||   |||
2254 nGluArgIleAsp...AlaTrpAlaGlnLeuAsnSerIle.......... 2266

1346 TCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACT 1395
           |||          :::|||||||:::              :::|||
2267 ........ProGlyGlnPheThrGlySerThrGlyValGlnValLeuThr 2280

1396 GCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATAT 1445
        :::            :::|||   :::   |||          :::||
2281 GlnGlu............GlnLeuAlaAsnThrGlyAlaGlnAlaTrpIl 2293

1446 CCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATA 1495
     |          ::::::      :::|||||||:::      |||   :::
2293 eLysLysAspGlnPheLeuArgAlaAlaProValThrGlyGlyMetGlyA 2310

1496 ATGGATATTTGGATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAGA 1545
           :::   |||:::|||      |||   :::|||   :::
2310 laValLeuMetArgLysMetGly......TrpArgGluGlyGluGlyLeu 2324

1546 ACTAAAGGTCAAGAA 1560
        |||:::::|||
2325 GlyLysAsnLysGlu 2329
```

```
seq_name: SwissProt_40:PHP_DROME

seq_documentation_block:
ID   PHP_DROME      STANDARD;      PRT;  1589 AA.
AC   P39769;
DT   01-FEB-1995 (Rel. 31, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Polyhomeotic-proximal chromatin protein.
GN   PH-P.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Imaginal disks;
RX   MEDLINE=92146957; PubMed=1346609;
RA   Decamillis M., Cheng N.S., Pierre D., Brock H.W.;
RT   "The polyhomeotic gene of Drosophila encodes a chromatin protein that
RT   shares polytene chromosome-binding sites with Polycomb.";
RL   Genes Dev. 6:223-232(1992).
RN   [2]
RP   SEQUENCE OF 199-1584 FROM N.A.
RX   MEDLINE=92039031; PubMed=1937015;
RA   Deatrick J., Daly M., Randsholt N.B., Brock H.W.;
RT   "The complex genetic locus polyhomeotic in Drosophila melanogaster
RT   potentially encodes two homologous zinc-finger proteins.";
RL   Gene 105:185-195(1991).
CC   -!- FUNCTION: BINDS TO POLYTENE CHROMOSOMES. SEEMS TO INTERACT WITH
CC       PC. MAY INTERACT WITH PROTEINS ALREADY BOUND TO PROMOTER
CC       COMPLEXES AND MAY BE A NEGATIVE REGULATOR OF HOMEOTIC AND
CC       SEGMENTATION GENES. PLAYS A ROLE IN REGULATING THE EXPRESSION OF
CC       OTHER PAIR-RULE GENES SUCH AS EVE, FTZ, AND H.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- TISSUE SPECIFICITY: SALIVARY GLANDS.
CC   -!- SIMILARITY: CONTAINS 1 SAM DOMAIN.
CC   -!- SIMILARITY: TO MOUSE EARLY DEVELOPMENT REGULATOR PROTEIN RAE-28.
CC   -!- CAUTION: IT IS UNCERTAIN WHETHER MET-1 OR MET-9 IS THE INITIATOR.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63672; CAA45211.1; -.
DR   EMBL; M64750; -; NOT_ANNOTATED_CDS.
DR   PIR; S23632; S23632.
DR   FlyBase; FBgn0004861; ph-p.
DR   InterPro; IPR001660; SAM.
DR   Pfam; PF00536; SAM; 1.
DR   SMART; SM00454; SAM; 1.
DR   PROSITE; PS50105; SAM_DOMAIN; 1.
KW   Zinc-finger; Developmental protein; DNA-binding; Nuclear protein.
FT   ZN_FING    1365   1387       C4-TYPE.
FT   DOMAIN     1513   1577       SAM.
FT   DOMAIN       74     80       POLY-GLN.
FT   DOMAIN      411    450       GLN-RICH.
FT   DOMAIN      494    520       GLN-RICH.
FT   DOMAIN      619    650       GLN-RICH.
FT   DOMAIN      775    960       GLN-RICH.
FT   DOMAIN     1233   1290       SER/THR-RICH.
FT   CONFLICT    254    254       MISSING (IN REF. 2).
FT   CONFLICT   1415   1415       D -> A (IN REF. 2).
SQ   SEQUENCE   1589 AA;  167297 MW;  A6DF0CF9106E1891 CRC64;

alignment_scores:
          Quality:  125.00              Length:    537
            Ratio:  0.551                 Gaps:     26
Percent Similarity:  42.272   Percent Identity: 21.974

alignment_block:
US-09-303-518D-465 x PHP_DROME   ..

Align seg 1/1  to: PHP_DROME  from: 1  to: 1589

     74 CAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGACCGTCAGCAT 123
        |||      |||      |||  |||      |||||||::  |||
    955 GlnSerGlyGlnLeuGlnLeuSerSerValProPheSerValSerSerSe 971

    124 TTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGGGGAACTTGC 173
        :::::::|||:::|||  :::|||  ||||||      :::  |||
    971 rThrThrProAlaGlyIleAlaThrSerSerAlaLeuGlnAlaAlaLeuS 988

    174 CGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGCCATCAGTTGG 223
          :::::::::::|||      ::::::|||  |||
    988 erAlaSerGlyAlaIle......PheGlnThrAlaLysPro..... 999

    224 GCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATT 273
        :::|||||||||||||||                    |||  |||
   1000 GlyThrCysSerSerSer.....................SerProThrSe 1009

    274 GTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCCC......CCTTCGA 317
        |||:::      |||||||:::  :::|||  |||        |||:
   1009 rSerValValThrIleThrAsnGlnSerSerThrProLeuValThrSerS 1026

    318 CAACCATG..................................... 325
        ::|||:::
   1026 erThrValAlaSerIleGlnGlnAlaGlnThrGlnSerAlaGlnValHis 1042

    325 ..... 325

   1043 GlnHisGlnGlnLeuIleSerAlaThrIleAlaGlyGlyThrGlnGlnGl 1059

    326 .............CCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGT 362
                    |||  :::      :::|||::::::  :
   1059 nProGlnGlyProProSerLeuThrProThrThrAsnProIleLeuAlaM 1076

    363 TGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATC 412
        ::|||  |||  ::::::|||:::::|||        :::|||
   1076 etThr...SerMetMetAsnAlaThrValGly...HisLeuSerThrAla 1090

    413 CCGCCG.....................ACGGCTATGACGGGCCACAGGGC 441
        ||||||                    ||||||:::|||:::      ::
   1091 ProProValThrValSerValThrSerThrAlaValThrSerSerProGl 1107

    442 GGCGGCTATCCCGCTCCCAAAGGCGCGA............GGGATATATA 479
        :      :::  |||      |||          |||
   1107 yGlnLeuValLeuLeuSerThrAlaSerSerGlyGlyGlySerIleP 1124

    480 CAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCG 529
          ||||||                                    |||
   1124 roAlaThr.....................................Pro 1127

    530 ACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGT 579
        |||     :::|||:::::|||      |||::::::  :::|||::
   1128 ThrLysGluThrProSerLysGly...ProThrAlaThrLeuValProIl 1143

    580 AGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATA 629
        :                                    :::  |||  |
   1143 e.................................GlySerProLysT 1148

    630 CAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCA 679
        ||      |||      |||  :::      ||||||  ||||||
   1148 hrProValSerGlyLysAspThrCysThrThrProLysSerSerThrPro 1164

    680 CTGCAGATATCGTCA......AAAACATCATCGGCGCGGCAGGAGAAATT 723
              :::|||:::      ::::::||||||  ::::::||||||
   1165 AlaThrValSerAlaSerValGluAlaSerSer..SerThrGlyGluAla 1180
```

```
 724 GTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGC 773
     ::::::   |||||||||   :::      |||      |||:::
1181 LeuSerAsnGlyAspAlaSerAspArgSerSerThrLeuSerLys..... 1195

 774 TGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCA 823
                  |||      :::      :::|||:::::::::
1196 ..............GlyAlaThrThrProThrSerLysGlnSerAsn. 1206

 824 TCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCC 873
                                                  ||||||
1207 ..........................................AlaAla 1208

 874 ATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAA.. 921
     ::::::      :::      :::||||||::::::::|||   |||
1209 ValGlnProProSerSerThrThrProAsnSerValSerGlyLysGluGl 1225

 922 ..............................GCCGTCAGCAATATCT 937
                                    |||:::|||:::
1225 uProLysLeuAlaThrCysGlySerLeuThrSerAlaThrSerThrSerT 1242

 938 TTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATAC 987
       |||:::::::|||   ::::::|||||||||:::::::|||
1242 hrThrThrThrIle...ThrAsnGlyIleGlyValAlaArgThrThrAla 1257

 988 GGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGA 1037
     :::  :::      ||||||      :::      |||
1258 SerThrAlaValSerThrAlaSerThrThrThrThrSer..... 1270

1038 GATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATG 1087
                                       |||   :::|||    , :
1271 ..............................SerGlyThrPheIleThrS 1277

1088 CGGCATACGCCAAATACCCGTCCCCTTACCATTCC...CGAAATATCCGT 1134
     :::::   :::          :::          |||      |||
1277 erCysThrSerThrThrThrThrThrThrSerSerIleSerAsnGlySer 1293

1135 TCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGT 1184
     :::::::|||   :::          |||   |||:::   :::   :::::
1294 LysAspLeuProLysAlaMetIleLysProAsnValLeuThrHisValIl 1310

1185 GCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGA 1234
     :          :::|||      ::::::   ||||||:::      |||
1310 e.........AspGlyPheIleIleGlnGluAlaAsnGluProPheProV 1324

1235 AGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGAC 1284
       |||:::      :::                          :::||||||
1324 alThrArgGlnArgTyr.....................AlaAspLysAsp 1333

1285 GTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTAT 1334
     |||:::
1334 ValSer..................................... 1335

1335 AGATGAACCCGTCTTTAATCCTAAA..................... 1359
      |||||||||         ||||||
1336 .AspGluPro.........ProLysLysLysAlaThrMetGlnGluAspI 1349

1360 ........GGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGA 1401
             :::::::::::::|||||||   :::   ::::::::::|||
1349 leLysLeuSerGlyIleAlaSerAlaProGlySerAspMetValAlaCys 1365

1402 ATTCAATACGCAAAATTACCAAGGCAAGGTAGAATC.........AGATA 1442
       |||   :::|||:::   ::::::::::::::         |||||
1366 GluGlnCysGlyLysMetGluHisLysAlaLysLeuLysArgLysArgTy 1382

1443 TATCCCACCT 1452
     |      |||
1382 rCysSerPro 1385

seq_name: SwissProt_40:A180_MOUSE

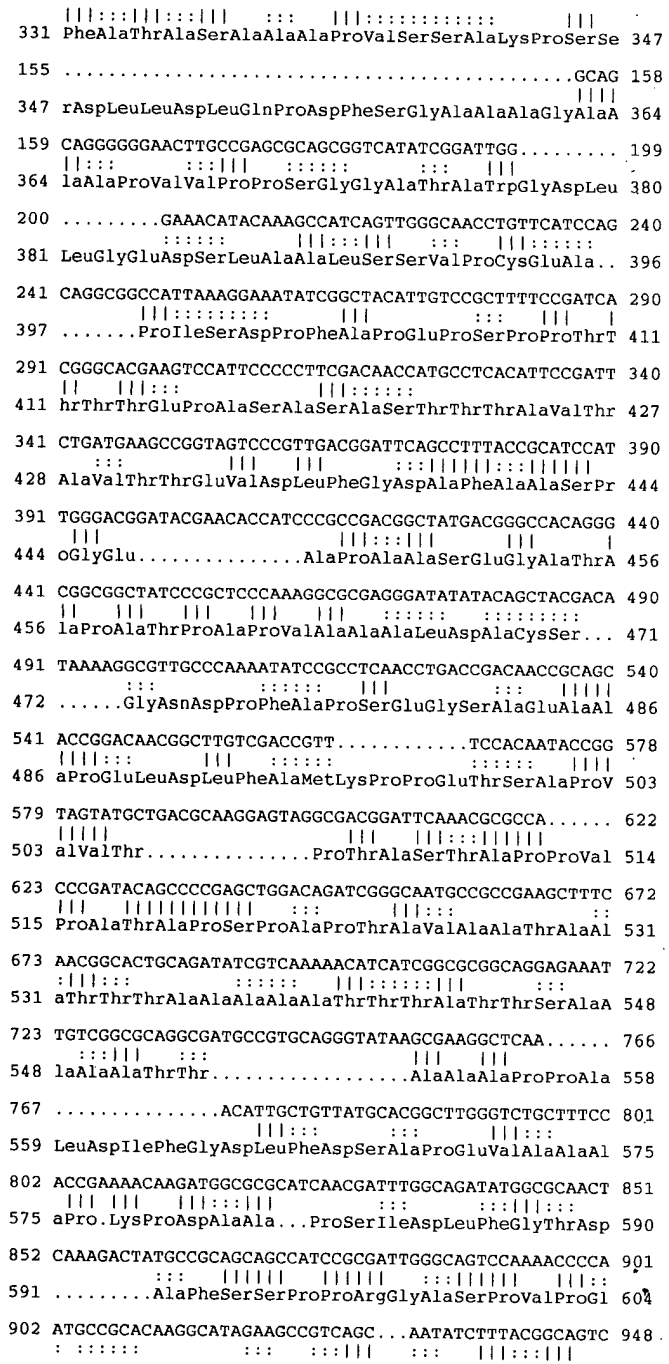
seq_documentation_block:
ID   A180_MOUSE     STANDARD;      PRT;   901 AA.
AC   Q61548; Q61547;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Clathrin coat assembly protein AP180 (Clathrin coat associated protein
DE   AP180) (Phosphoprotein F1-20) (91 kDa synaptosomal-associated
DE   protein).
GN   SNAP91.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92300439; PubMed=1607933;
RA   Zhou S., Sousa R., Tannery N.H., Lafer E.M.;
RT   "Characterization of a novel synapse-specific protein. II. cDNA
RT   cloning and sequence analysis of the F1-20 protein.";
RL   J. Neurosci. 12:2144-2155(1992).
CC   -!- FUNCTION: ADAPTINS ARE COMPONENTS OF THE ADAPTOR COMPLEXES WHICH
CC       LINK CLATHRIN TO RECEPTORS IN COATED VESICLES. CLATHRIN-
CC       ASSOCIATED PROTEIN COMPLEXES ARE BELIEVED TO INTERACT WITH THE
CC       CYTOPLASMIC TAILS OF MEMBRANE PROTEINS, LEADING TO THEIR SELECTION
CC       AND CONCENTRATION. BINDING OF AP180 TO CLATHRIN TRISKELIA INDUCES
CC       THEIR ASSEMBLY INTO 60-70 NM COATS.
CC   -!- SUBCELLULAR LOCATION: COMPONENT OF THE COAT SURROUNDING THE
CC       CYTOPLASMIC FACE OF COATED VESICLES IN THE PLASMA MEMBRANE.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; A LONG FORM (SHOWN HERE) AND A
CC       SHORT FORM; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: BRAIN. ASSOCIATED WITH THE SYNAPSES.
CC   -!- DEVELOPMENTAL STAGE: DEVELOPMENTALLY REGULATED IN A PATTERN
CC       COINCIDENT WITH ACTIVE SYNAPTOGENESIS AND SYNAPTIC MATURATION.
CC   -!- DOMAIN: POSSESSES A THREE DOMAIN STRUCTURE: THE N-TERMINAL 300
CC       RESIDUES HARBOR A CLATHRIN BINDING SITE, AN ACIDIC MIDDLE DOMAIN
CC       450 RESIDUES, INTERRUPTED BY AN ALA-RICH SEGMENT, AND THE C-
CC       TERMINAL DOMAIN (166 RESIDUES).
CC   -!- PTM: PHOSPHORYLATED.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; M83985; AAA37587.1; -.
DR   EMBL; M83985; AAA37586.1; -.
DR   HSSP; P04002; 1WFA.
DR   MGD; MGI:109132; Snap91.
DR   InterPro; IPR001026; ENTH.
DR   Pfam; PF01417; ENTH; 1.
DR   SMART; SM00273; ENTH; 1.
KW   Coated pits; Alternative splicing; Phosphorylation.
FT   DOMAIN      410    413       POLY-THR.
FT   DOMAIN      535    539       POLY-ALA.
FT   DOMAIN      547    550       POLY-ALA.
FT   DOMAIN      659    664       POLY-SER.
FT   DOMAIN      704    710       POLY-SER.
FT   VARSPLIC    715    719       MISSING (IN SHORT ISOFORM).
SQ   SEQUENCE   901 AA;  91851 MW;  24A98FBACE8DB8B1 CRC64;


alignment_scores:
            Quality:  124.00              Length:  546
              Ratio:  0.515                 Gaps:   27
 Percent Similarity:  44.139   Percent Identity: 21.795

alignment_block:
 US-09-303-518D-465 x A180_MOUSE   ..

 Align seg 1/1  to: A180_MOUSE  from: 1  to: 901

     107 TTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCG.. 154
```



seq_name: SwissProt_40:YM96_YEAST

seq_documentation_block:
ID YM96_YEAST STANDARD; PRT; 1140 AA.
AC Q04893;
DT 01-NOV-1997 (Rel. 35, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 01-NOV-1997 (Rel. 35, Last annotation update)
DE Hypothetical 113.1 kDa protein in PRE5-FET4 intergenic region.
GN YMR317W OR YM9924.09.
OS Saccharomyces cerevisiae (Baker's yeast).
OC Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX NCBI_TaxID=4932;
RN [1]

```
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Churcher C.M., Louis E.J., Barrell B.G., Rajandream M.A., Walsh S.V.;
RL   Submitted (NOV-1995) to the EMBL/GenBank/DDBJ databases.
CC   -!- DOMAIN: CONTAINS MANY SER/THR-RICH DOMAIN AND REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z54141; CAA90835.1; -.
DR   SGD; S0004936; YMR317W.
KW   Hypothetical protein; Repeat.
SQ   SEQUENCE   1140 AA;  113070 MW;  0153EBCA24FE5427 CRC64;


alignment_scores:
            Quality:  123.50              Length:    423
              Ratio:  0.585                Gaps:     18
 Percent Similarity:  49.882   Percent Identity: 22.695

alignment_block:
 US-09-303-518D-465 x YM96_YEAST   ..

 Align seg 1/1  to: YM96_YEAST  from: 1  to: 1140

     110 TCGACCGTCAGCATTTCGAACCCGACGGGAAATACC............AC 147
         ||||||:::|||   :::   |||   :::::::||            ||
     240 SerThrIleSerGluThrLeuProPheSerSerThrIleLeuSerIleTh 256

     148 CTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATT 197
         |   |||   ::::::::::::   ||||||   |||:::   |||:::
     256 rSerSerProValSerSerGluAlaProSer...AlaThrSerSerServ 272

     198 GGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGG 247
           ::::::     :::   |||   ::::::      |||||||:::
     272 alSerSerGluAlaSerSerThrSerSerServalSerSerGluAla 288

     248 CCATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCAC 297
         ||||||       |||::::::   |||     |||   :::::::||
     289 ProLeuAlaThrSerServalValSerSerGluAlaPro...SerSerTh 304

     298 GAAGTCCATT............CCCCCTTCGACAACCATGCCTCACATTC 335
         |:::|||:::              ||||||:::|||         :::
     304 rSerSerValValSerSerGluAlaProSerSerThrSerSerSerVal. 320

     336 CGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCA 385
                                        :::|||   :::::::::
     321 ..............................SerSerGluIleSerSer 326

     386 TCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCA 435
         :::   ::::::::::::::::::      |||   |||   ::::::
     327 ThrThrSerSerSerValSerSerGluAlaProLeuAlaThrSerSerVa 343

     436 CAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATA.......... 475
             ::::::      |||         :::   ::::::
     343 lValSerSerGluAlaProSerSerThrSerSerSerValSerSerGluI 360

     476 ........TATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCC 517
                 |||:::::::::   ::::::         :::::::::
     360 leSerSerThrThrSerSerSerServalSerSerGluAlaProLeuAlaThr 376

     518 TCAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTC 567
         |||:::      ::::::   ||||||::::::      |||:::|||||
     377 SerSerValValSerSerGluAlaProSerSerThrSerSerSerValSe 393

     568 CACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACG 617
         |                                           |||:::|
     393 r.........................................SerGluA 396

     618 CGCCACCCGATACA.....................GCCCCGAGCTGGACA 646
         |||||   :::|||                     |||||||||   |||
     396 laProSerSerThrSerSerSerValSerSerGluAlaProSerSerThr 412

     647 GATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAG......ATATC 690
         :::   ::::::      :::::::|||::::::   :::       :::||
     413 SerSerSerValSerSerGluIleSerSerThrLysSerSerValMetSe 429

     691 GTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGC 740
         ||||:::::|||||||||   ::::::   :::|||::::::|||   |
     429 rSerGluValSerSerAlaThrSerSerLeuValSerSerGluAla...P 445

     741 CGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGG 790
         ||      :::   :::   |||::::::   |||
     445 roSerAlaIleSerSerLeuAlaSerSerArgLeu.............. 456

     791 GTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGAT 840
               |||      |||:::         :::|||||||:::
     457 ......PheSerSerLysAsnThrSerValThrSerThrLeuValAlaTh 471

     841 ATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGT 890
               :::|||  |||         :::||||||          |
     471 rGluAlaSerSerValThrSerSerLeuArgProSer...........S 484

     891 CCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATCTTTA 940
         ||:::|||                          ::::::   |||:::
     484 erGluThr.....................LeuAlaSerAsnSerIle 492

     941 CGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGC 990
            :::||||||   |||   |||   :::         :::::|||::
     493 IleGluSerSerLeuSerThrGlyTyrAsnSerThrValSerThrThrTh 509

     991 TTGGGCGGCATCA.............CGGCACATCCTGTCAAGCGGTCGCA 1028
         :   |||||||||                  :::   |||||||:::   |
     509 rSerAlaAlaSerSerThrLeuGlySerLysValSerSerSerAsnSerA 526

    1029 GATGGGCGAGATCGCATTGCCGAAAG....................... 1054
         ||   |||   |||        :::
     526 rgMetAlaThrSerLysThrSerSerThrSerSerAspLeuSerLysSer 542

    1055 ............GGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCA 1092
                     ||||||      ::::::|||      |||
     543 SerValIlePheGlyAsnSerSerThrValThrThrSerProSerAlaSe 559

    1093 TACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTT 1142
                   |||            :::|||::::::            |
     559 rIleSerLeuThrAlaSer...ProLeuProSerVal............T 571

    1143 GGAGCAGCGTTACGGCAAA..................GAAAACATCACC 1173
         |||||::::::|||:::                    :::|||::::::
     571 rpSerAspIleThrSerSerGluAlaSerSerIleSerSerAsnLeuAla 587

    1174 TCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAA 1223
         ||||||::::::|||   :::|||:::::::::::::      |||:::
     588 SerSerSerAlaProSerAspAsnAsnSerThrIleAlaSerAlaSerLe 604

    1224 ACGCCACCCGAAGACCAAA 1242
                |||||||||
     604 uIleValThrLysThrLys 610

seq_name: SwissProt_40:VGP3_EBV

seq_documentation_block:
ID   VGP3_EBV       STANDARD;      PRT;   907 AA.
AC   P03200; P03201;
DT   21-JUL-1986 (Rel. 01, Created)
DT   21-JUL-1986 (Rel. 01, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Envelope glycoprotein GP340 (Membrane antigen) (MA) [Contains:
DE   Glycoprotein GP220].
GN   BLLF1.
OS   Epstein-barr virus (strain B95-8) (Human herpesvirus 4).
```

```
OC   Viruses; dsDNA viruses, no RNA stage; Herpesviridae;
OC   Gammaherpesvirinae; Lymphocryptovirus.
OX   NCBI_TaxID=10377;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=84270667; PubMed=6087149;
RA   Baer R., Bankier A.T., Biggin M.D., Deininger P.L., Farrell P.J.,
RA   Gibson T.J., Hatfull G., Hudson G.S., Satchwell S.C., Seguin C.,
RA   Tuffnell P.S., Barrell B.G.;
RT   "DNA sequence and expression of the B95-8 Epstein-Barr virus genome.";
RL   Nature 310:207-211(1984).
CC   -!- FUNCTION: RESPONSIBLE FOR EBV BINDING TO THE CR2 RECEPTOR ON HUMAN
CC       B-CELLS.
CC   -!- SUBCELLULAR LOCATION: MOST ABUNDANT COMPONENT OF THE VIRAL
CC       ENVELOPE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; V01555; CAA24854.1; -.
DR   PIR; A03762; QQBE21.
DR   PIR; A03763; QQBE22.
DR   PIR; S33008; S33008.
KW   Membrane; Glycoprotein; Antigen; Late protein; Alternative splicing.
FT   CARBOHYD     47     47       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     87     87       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    114    114       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    166    166       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    169    169       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    195    195       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    229    229       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    277    277       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    318    318       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    328    328       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    345    345       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    356    356       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    378    378       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    386    386       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    411    411       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    435    435       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    443    443       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    457    457       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    497    497       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    519    519       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    533    533       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    547    547       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    568    568       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    589    589       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    610    610       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    624    624       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    627    627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    645    645       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    656    656       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    683    683       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    701    701       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    735    735       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    746    746       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    755    755       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    780    780       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    815    815       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    858    858       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    888    888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC    502    698       MISSING (IN GP220).
SQ   SEQUENCE   907 AA;  94431 MW;  0750141CBCAC52C9 CRC64;

alignment_scores:
           Quality:  123.00              Length:     566
               Ratio:  0.498                 Gaps:     31
 Percent Similarity:  43.640   Percent Identity: 23.322

alignment_block:
 US-09-303-518D-465 x VGP3_EBV   ..

 Align seg 1/1  to: VGP3_EBV  from: 1  to: 907

      95 TTATCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATAC 144
         :::||||||   |||:::           ||||||   |||   :::
     379 IleSerGlyAlaPheAla.....SerAsnArgThrPheAspIl 391

     145 CACCTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGG 194
          |||   |||:::      |||:::   |||::::::   :::|||:::
     391 eThrValSerGlyLeu...GlyThrAlaProLysThrLeuIleIleThrA 407

     195 ATTGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGG 244
               |||      :::   :::                    |||:::
     407 rgThrAlaThrAsnAlaThrThrThrThrHisLysValIlePheSerLys 423

     245 CGGCCATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGG 294
          |||      |||   ::::::|||      ::::::      |||||
     424 AlaPro.....GluSerThrThrThrSerProThrLeuAsnThrThrGl 438

     295 C.................ACGAAGTCCATTCCCCCTTCGACAACCATGC 326
         |                 |||   ::::::|||   ||||||
     438 yPheAlaAspProAsnThrThrThrGlyLeuProSerSerThr..... 452

     327 CTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCC 376
           |||:::|||                          |||:::   :::
     453 ..HisValPro.......................ThrAsnLeuThr 459

     377 TTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTAT 426
              ||||||   |||      :::::::|||          |||:::
     460 AlaProAlaSerThrGlyProThrValSerThrAlaAspValThrSerPr 476

     427 GACG......GGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAG 470
          |||      |||      :::::::|||   |||:::      :::   :
     476 oThrProAlaGlyThrThrSerGlyAlaSerProValThrProSerProS 493

     471 GGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCA 520
         ::   :::          |||   :::      |||   :::::::::|||
     493 erProTrpAspAsnGlyThrGluSerLysAlaProAspMetThrSerSer 509

     521 ACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCAC 570
         |||   |||:::|||:::          |||          :::::::|||
     510 ThrSerProValThrThrProThrProAsn.........AlaThrSerPr 523

     571 AATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGC 620
           |||:::|||                 ||||||   |||   :::|
     523 oThrProAlaValThrThrProThrProAsnAlaThr...SerProThrP 539

     621 CACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCG...... 664
         ||      |||:::|||      |||   :::|||      |||
     539 roAlaValThrThrPro......ThrProAsnAlaThrSerProThrLeu 553

     665 ...AAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCG 711
            |||   |||   :::      :::::::::   |||   :::|||
     554 GlyLysThrSerProThrSerAlaValThrThrProThrProAsnAlaTh 570

     712 GCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGG 761
          :::      |||:::      :::           |||   :::   :
     570 rSerProThrLeuGlyLysThrSerProThrSerAlaValThrThrProT 587

     762 CTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCA......... 802
         ::   :::          |||               |||
     587 hrProAsnAlaThrSerProThrLeuGlyLysThrSerProThrSerAla 603

     803 .........CCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATG 843
                 |||      :::      :::   |||:::
     604 ValThrThrProThrProAsnAlaThrGlyProThrVal........... 616

     844 GCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATT.......... 883
```

```
           ::::::|||   ||||||      ::::::
  617 .......GlyGluThrSerProGlnAlaAsnAlaThrAsnHisThrLeuG 631

  884 ..GGGCAGTCCAAAACCCCAATG...........CCGCACAAGGCATAG 919
        |||   |||   ||||||:::            |||   :::|||
  631 lyGlyThrSerProThrProValValThrSerGlnProLysAsnAla... 646

  920 AAGCCGTCAGCAATATCTTTACGGCAGTCATCC................. 952
          ||||||::::::      |||   :::
  647 ...ThrSerAlaValThrThrGlyGlnHisAsnIleThrSerSerSerTh 662

  953 ..................CCGTCAAAGGGATTGGAGCTGTTCGGGGAAA 983
                       ||||||::::::   |||   ::::::   :
  662 rSerSerMetSerLeuArgProSerSerAsnProGluThrLeuSerProS 679

  984 ATACGGCTTGGGCGGCATCACGGCAC......ATCCTGTCAAGCGGTCGC 1027
      ::|||:::      ::::::   |||      :::|||:::|||:::::
  679 erThrSerAspAsnSerThrSerHisMetProLeuLeuThrSerAlaHis 695

 1028 AGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCA...GCGAC 1074
          :::   ::::::      :::   :::|||   |||   :::||
  696 ProThrGlyGlyGluAsnIleThrGlnValThrProAlaSerIleSerTh 712

 1075 AATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCCCG 1124
      |              :::|||      ||||||   |||      :
  712 rHisHisValSerThrSerSerProAlaProArgProGlyThrThrSerG 729

 1125 AAATATCCGT.........TCAAACTTGGAGCAGCGTTACGGCAAAGAA 1164
      ::   |||           |||:::   :::   :::   |||:::
  729 lnAlaSerGlyProGlyAsnSerSerThrSerThrLysProGlyGluVal 745

 1165 AACATCACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACT 1214
      |||:::|||:::::|||   |||||||:::|||:::   :::   :::
  746 AsnValThrLysGlyThr...ProProGlnAsnAlaThrSerProGlnAl 761

 1215 GGCAAACAAACGCCACCCGAAGACCAAAGTGCCG......TTTGACGGTA 1258
          :::   :::       ||||||   ||||||        :::
  761 aProSerGlyGln......LysThrAlaValProThrValThrSerThrG 776

 1259 AAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACG......... 1299
        |||      |||             |||:::   |||
  776 lyGlyLysAlaAsnSerThrThrGlyGlyLysHisThrThrGlyHisGly 792

 1300 ...AGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGT 1346
         |||   :::|||                                ||||||::
  793 AlaArgThrSerThr.........................GluProTh 800

 1347 CTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTG 1396
      :   :::      |||:::                          |||   |||
  800 rThrAspTyrGlyGlyAsp.....................SerThrThrP 810

 1397 CCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATC 1446
        |||                 ||||||   :::            |||:::
  810 roArg.................ProArgTyrAsnAlaThrThrTyrLeu 820

 1447 CCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAA 1496
      ||||||::::::   |||         |||
  821 ProProSerThrSerSerLysLeuArgPro..................... 830

 1497 TGGATATTTGGATAAATTTGGTAATGAATGGACT...AAAGGTCCATCAA 1543
                                 :::||||||      :::|||
  831 .......................ArgTrpThrPheThrSerProProV 839

 1544 GAACTAAAGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACAGGA 1593
        |||   :::|||            |||   :::      |||:::
  839 alThrThrAlaGlnAla.........ThrValProValProProThrSer 852

 1594 AGAGAGCAA.....................CTTGGATGGGCTAGT 1617
      :::   :::                    |||   |||||||||
  853 GlnProArgPheSerAsnLeuSerMetLeuValLeuGlnTrpAlaSer 868

seq_name: SwissProt_40:PODX_RABIT
```

```
seq_documentation_block:
ID   PODX_RABIT     STANDARD;      PRT;   551 AA.
AC   Q28645;
DT   01-MAR-2002 (Rel. 41, Created)
DT   01-MAR-2002 (Rel. 41, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Podocalyxin-like protein 1 precursor.
GN   PODXL OR PCLP1.
OS   Oryctolagus cuniculus (Rabbit).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Lagomorpha; Leporidae; Oryctolagus.
OX   NCBI_TaxID=9986;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NEW ZEALAND WHITE;
RX   MEDLINE=96094343; PubMed=7493982;
RA   Kershaw D.B., Thomas P.E., Wharram B.L., Goyal M., Wiggins J.E.,
RA   Whiteside C.I., Wiggins R.C.;
RT   "Molecular cloning, expression, and characterization of podocalyxin-
RT   like protein 1 from rabbit as a transmembrane protein of glomerular
RT   podocytes and vascular endothelium.";
RL   J. Biol. Chem. 270:29439-29446(1995).
CC   -!- FUNCTION: Functions as an antiadhesin that maintains an open
CC       filtration pathway between neighboring foot processes in the
CC       podocyte by charge repulsion.
CC   -!- SUBCELLULAR LOCATION: Type I membrane protein (Potential).
CC   -!- TISSUE SPECIFICITY: Glomerular epithelium cell (podocyte).
CC   -!- PTM: Sialoglycoprotein.
CC   -!- SIMILARITY: BELONGS TO THE PODOCALYXIN FAMILY.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; U35239; AAC48489.1; -.
KW   Glycoprotein; Signal; Transmembrane.
FT   SIGNAL        1     21       POTENTIAL.
FT   CHAIN        22    551       PODOCALYXIN-LIKE PROTEIN 1.
FT   DOMAIN       22    452       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    453    473       POTENTIAL.
FT   DOMAIN      474    551       CYTOPLASMIC (POTENTIAL).
FT   CARBOHYD    145    145       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    180    180       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    333    333       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   551 AA;  57040 MW;  E9B8AE168CDFB8C5 CRC64;


alignment_scores:
            Quality:  121.50              Length:    385
              Ratio:  0.719                 Gaps:     20
 Percent Similarity:  43.896   Percent Identity: 24.675

alignment_block:
 US-09-303-518D-465 x PODX_RABIT   ..

 Align seg 1/1  to: PODX_RABIT  from: 1  to: 551

     272 TTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAAC 321
         ||||||         ::::::      ||||||      |||:::   ||
      15 LeuSerProProSerLeuSerGlnGluLysSerProGlnProGlyProTh 31

     322 CATGCCTCACATTCCGATTCTGATGAAGCCGGTA.........GTCCCGT 362
         |                              :::|||:::         :::|||
      31 rProMetAlaThrSerThrSerThrArgProAlaProAlaSerAlaProA 48

     363 TGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATC 412
                 |||:::   :::|||||||:::   ::::::   ||||||
      48 laProLysSerSerValAlaAlaSerValProAlaGluGlnAsnThrThr 64
```

```
 413 CCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAA 462
     |||   |||:::             :::|||   |||         |||::
  65 ProMetThrThr...........LysAlaProAlaThrGlnSerProSe 77

 463 GGCGCGA..................................................... 469
     :|||
  77 rAlaSerProGlySerSerValGluAsnSerAlaProAlaGlnGlySerT 94

 469 ..................................................... 469

  94 hrThrThrGlnGlnSerLeuSerValThrThrLysAlaGluAlaLysAsp 110

 470 ......GGGATATATACAGCT...ACGACATAAAAGGCGTTGCCC..... 505
           |||:::   ||||||   :::|||   :::|||   |||
 111 AlaGlyGlyValProThrAlaHisValThrGlySerAlaArgProValTh 127

 506 ..........AAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGG 545
             :::::::::|||:::         ::::::   ||||||:
 127 rSerGlySerGlnValAlaAlaGlnAspProAlaAlaSerLysAlaProS 144

 546 ACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAG 595
     ::|||         :::::::::|||   |||::::::
 144 erAsn......HisSerIleThrThrLysProLeuAlaThrGluAlaThr 158

 596 GAGTAGGCG............ACGGATTCAAACGCGCCACCCGATACAGC 633
     :::   |||           ||||||   :::::::|||      |||||
 159 SerGlnAlaProArgGlnThrThrAspValGlyThrProGlyProThrAl 175

 634 CCCGAGCTGGACAGATCGGGCAATGCCG.....................C 662
     ||||      |||:::   :::   |||                      |
 175 aProProValThrAsnSerThrSerProAspLeuLeuGlyHisAlaThrP 192

 663 CGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATC.GGCGCG 711
     |||||   |||   :::   |||:::|||      |||:::   |||:::
 192 roLysProSerGluGlyProGlnLeuSerPheProThrAlaAlaGlySer 208

 712 GCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGG 761
        |||   ::::::|||:::|||
 209 LeuGlyProValThrGlySerGlyThr..................... 217

 762 CTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACC...GAAA 808
                         |||   |||   |||||||||   ::::
 218 .....................GlySerGlyThrLeuSerThrProGlnG 227

 809 ACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTCAAAGAC 858
     ::|||   |||   ::::::   :::|||:::   |||::: ::::::
 227 lyLysProAlaThrLeuThrProValAlaSerSerAlaGlu.ThrGlnGl 243

 859 TATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGC 908
      ||||||                          :::||||||||||
 243 yMetPro............................SerProMetProP 250

 909 ACAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCA...TCCCCG 955
       :::      :::|||||||:::      :::   :::|||   ||||||
 250 roSerProAlaSerProSerSerSerProProheProSerSerProSerPro 266

 956 TCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCACG 1005
     |||   :::|||:::      |||   :::::::|||   :::   :::
 267 SerProAlaLeuGlnProSerGlyProSerAlaAlaGlyThrGluAsp.. 282

1006 GCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGG 1055
              ::::::||||||                           |
 283 ..........ThrThrGlyArg.............................G 287

1056 GAAATCCGCCGTCAGCGACAATTTTGCCGATGCGG......CATACGCCA 1099
     ||         |||:::|||   |||                |||   |||
 287 lyProThrSerSerSerThrGluLeuAlaSerThrAlaLeuHisGlyPro 303

1100 AATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAG 1149
     :::|||         |||                            |||::
 304 SerThrLeuSerProThr.........................SerAl 311

1150 CGTTACGGCAAAGAAAACATCA......CCTCCTCAACCGTGCCGCCGTC 1193
```

```
          :|||      :::::::::|||      ||||||:::        |||
  311 aValArgAspGlnArgValSerCysGlyProProGlu.........ArgP 325

 1194 AAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAG 1243
        ||||||::::::              |||   ||||||
  325 roThrGluGlnLeu......LeuIleLeuAsnLeuThrArgSerSerPro 339

 1244 TGC 1246
      |||
  340 Cys 340
```

```
seq_name: SwissProt_40:EGR1_BRARE

seq_documentation_block:
ID   EGR1_BRARE     STANDARD;      PRT;   511 AA.
AC   P26632;
DT   01-AUG-1992 (Rel. 23, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Early growth response protein 1 (EGR-1) (Krox24).
GN   EGR1.
OS   Brachydanio rerio (Zebrafish) (Zebra danio).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Actinopterygii; Neopterygii; Teleostei; Euteleostei; Ostariophysi;
OC   Cypriniformes; Cyprinidae; Danio.
OX   NCBI_TaxID=7955;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95032735; PubMed=7945937;
RA   Drummond I.A., Rohwer-Nutter P., Sukhatme V.P.;
RT   "The zebrafish egr1 gene encodes a highly conserved, zinc-finger
RT   transcriptional regulator.";
RL   DNA Cell Biol. 13:1047-1055(1994).
RN   [2]
RP   SEQUENCE OF 315-376 FROM N.A.
RX   MEDLINE=92028854; PubMed=1930167;
RA   Lanfear J., Jowett T., Holland P.W.;
RT   "Cloning of fish zinc-finger genes related to Krox-20 and Krox-24.";
RL   Biochem. Biophys. Res. Commun. 179:1220-1224(1991).
CC   -!- FUNCTION: TRANSCRIPTIONAL REGULATOR. RECOGNIZE AND BINDS TO THE
CC       DNA SEQUENCE 5'-CGCCCCCGC-3'(EGR-SITE). ACTIVATE THE TRANSCRIPTION
CC       OF TARGET GENES WHOSE PRODUCTS ARE REQUIRED FOR MITOGENESIS AND
CC       DIFFERENTIATION.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- INDUCTION: BY GROWTH FACTORS.
CC   -!- SIMILARITY: BELONGS TO THE EGR FAMILY OF C2H2-TYPE ZINC-FINGER
CC       PROTEINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U12895; AAA63651.1; -.
DR   EMBL; M81109; AAA50035.1; -.
DR   PIR; PQ0233; PQ0233.
DR   HSSP; P08046; 1AAY.
DR   ZFIN; ZDB-GENE-980526-320; egr1.
DR   InterPro; IPR000822; Znf-C2H2.
DR   Pfam; PF00096; zf-C2H2; 3.
DR   PRINTS; PR00048; ZINCFINGER.
DR   SMART; SM00355; ZnF_C2H2; 3.
DR   PROSITE; PS00028; ZINC_FINGER_C2H2_1; 3.
DR   PROSITE; PS50157; ZINC_FINGER_C2H2_2; 3.
KW   Transcription regulation; Activator; DNA-binding; Nuclear protein;
KW   Repeat; Zinc-finger; Metal-binding.
FT   DOMAIN      146    152       POLY-SER.
FT   DOMAIN      155    165       POLY-SER.
FT   DOMAIN      311    391       ZINC FINGERS.
FT   ZN_FING     311    335       C2H2-TYPE.
```

```
FT   ZN_FING     341    363       C2H2-TYPE.
FT   ZN_FING     369    391       C2H2-TYPE.
FT   CONFLICT    317    317       T -> S (IN REF. 2).
FT   CONFLICT    320    320       R -> S (IN REF. 2).
FT   CONFLICT    351    351       S -> R (IN REF. 2).
FT   CONFLICT    372    372       E -> D (IN REF. 2).
SQ   SEQUENCE   511 AA;  55139 MW;  411C31E6F4AA10BF CRC64;


alignment_scores:
            Quality:  121.00              Length:    424
              Ratio:  0.608                Gaps:     22
 Percent Similarity:  46.934   Percent Identity: 23.349

alignment_block:
 US-09-303-518D-465 x EGR1_BRARE   ..

 Align seg 1/1  to: EGR1_BRARE  from: 1  to: 511

      15 CAAAATATCCCTTATTCTGTCCATACTGGCAGTGTG.........CCTGC 55
         |||::::::|||          :::||||||                |||||
      90 GlnArgLeuProProIleSerTyrThrGlyArgPheThrLeuGluProAl 106

      56 CGATGCATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAG 105
         |      |||::::::  ::::::|||         |||:::|||
     106 aThrAsnCysSerAsn..SerLeuTrpAlaGluProLeuPheSerLeuVa 122

     106 GTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGG 155
           :::   |||:::|||:::   |||   ::::::          |||:
     122 lSerGlyLeuValGlyIleAsnProProProAlaSerIleProSerSerT 139

     156 CAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACA 205
         :::::   ::::::   ||||||:::::      |||:::      :::
     139 hrSerGlnAlaThrHisProSerSerSerSerThrSerSerIleProSer 155

     206 TACAAAGCCATCAGTTGGGCAACCTGTTCA..................... 235
           ::::::   |||   :::::::::|||
     156 SerSerSerSerSerThrSerSerAlaSerLeuSerCysSerValHisGl 172

     236 .TCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATTGTCCGCT.... 280
          |||:::   :::|||:::   :::   :::   |||   |||:::
     172 nSerGluProAsnProIleTyrSerAlaAlaProThrTyrSerSerAlaS 189

     281 ..........TTTCCGATCACGGGC........................ 295
                 ||||||   :::|||
     189 erProAspIlePheProGluSerGlyProAsnPheSerThrThrValGly 205

     296 ................................ACGAAGTCCATTCC 309
                                         :::|||:::
     206 ThrSerLeuGlnTyrSerSerSerThrTyrProSerAlaLysThrCysAs 222

     310 CCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCC 359
          ||||||   ::::::|||   ||||||
     222 nProSerPheSerValProMetIlePro..................... 231

     360 CGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACC 409
               |||      ||||||   :::      ::::::   :::
     232 ........AspTyrLeuPheThrGlnGln...GlnSerGluIleSerLeu 244

     410 ATCCCGCCG..............ACGGCTATGACGGGCCACAGGGCGGC 444
         :::||||||                :::::::::|||::::::   ||
     245 ValProProAspGlnLysProIleGlnThrGlnAlaGlyGlnGlnProAl 261

     445 GGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAA 494
         |      ||||||         :::::::::   |||   |||    :::
     261 aLeuThrProLeuHisThrIleLysAlaPheAlaThrGlnThr.GlySer 277

     495 AGG.....................................CGTTG 502
         :::                                     |||
     278 GlnAspLeuLysSerValTyrGlnSerGlnLeuIleLysProSerArgMe 294

     503 CCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAACGG 552
           |||||||||   :::|||::::::   |||   |||:::

     294 tArgLysTyrProAsnArgProSerLysThrProProHisGlu..... 308

     553 CTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGG...AGT 599
           ||||||:::                  :::|||   |||   |||
     309 .....ArgProTyrAlaCysProValGluThrCysAspArgArgPheSer 323

     600 AGGCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGAT 649
         |||           |||||||||   |||:::   ::::::||||||
     324 ArgSerAspGluLeuThrArgHis...IleArgIleHisThrGlyGln.. 338

     650 CGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAAC 699
                   :::   ||||||      ||||||      :::::::
     339 .............LysProPheGln......CysArgIleCysMetArgA 349

     700 ATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGG 749
         ::      |||:::   :::         :::   |||   ::::::|||
     349 snPheSerArgSerAspHisLeuThrThrHisIleArgThrHisThrGly 365

     750 TATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTT 799
           |||          :::   |||      :::::::::   :::
     366 GluLysProPheAlaCysGluIleCysGlyArgLysPheAlaArgSerAs 382

     800 CCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAA 849
          :::|||||| :::   ::::::|||:::            |||   |||
     382 pGluArgLys.ArgHisThrLysIleHis...........MetArgGln 394

     850 CTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCC 899
           ||||||      |||                |||||||||:::
     395 ...LysAspLysLysAlaGluLysGlyAlaThrAlaAlaValGlnSer.. 409

     900 CAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATC...TTTACGGCAG 946
                              :::||||||||||||   :::::::|||
     410 ......................SerValSerAsnIleSerIleSerAlaS 419

     947 TCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGC 996
           ||||||::::::      :::
     419 erSerProValSerSerTyrProSer....................... 427

     997 GGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATT 1046
           ||||||:::::|||         |||::::::
     428 ProIleThrSerTyrProSerProValSerSerPhe.............. 439

    1047 GCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACG 1096
                    |||   |||::::::   :::::::::   :::::::
     440 ..........ProSerProValAsnSerCysTyrSerSerProValHisT 453

    1097 CCAAATACCCGTCCCCT 1113
         :::::||||||||||||
     453 hrSerTyrProSerPro 458

seq_name: SwissProt_40:HAP1_HAEIN

seq_documentation_block:
ID   HAP1_HAEIN     STANDARD;      PRT;  1409 AA.
AC   P44596;
DT   01-NOV-1995 (Rel. 32, Created)
DT   01-NOV-1995 (Rel. 32, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Adhesion and penetration protein precursor (EC 3.4.21.-).
GN   HAP OR HI0248.
OS   Haemophilus influenzae.
OC   Bacteria; Proteobacteria; gamma subdivision; Pasteurellaceae;
OC   Haemophilus.
OX   NCBI_TaxID=727;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=RD / KW20 / ATCC 51907;
RX   MEDLINE=95350630; PubMed=7542800;
RA   Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F.,
RA   Kerlavage A.R., Bult C.J., Tomb J.-F., Dougherty B.A., Merrick J.M.,
RA   McKenney K., Sutton G., Fitzhugh W., Fields C.A., Gocayne J.D.,
RA   Scott J.D., Shirley R., Liu L.-I., Glodek A., Kelley J.M.,
RA   Weidman J.F., Phillips C.A., Spriggs T., Hedblom E., Cotton M.D.,
```

```
RA   Utterback T.R., Hanna M.C., Nguyen D.T., Saudek D.M., Brandon R.C.,
RA   Fine L.D., Fritchman J.L., Fuhrmann J.L., Geoghagen N.S.M.,
RA   Gnehm C.L., McDonald L.A., Small K.V., Fraser C.M., Smith H.O.,
RA   Venter J.C.;
RT   "Whole-genome random sequencing and assembly of Haemophilus
RT   influenzae Rd.";
RL   Science 269:496-512(1995).
CC   -!- FUNCTION: PROBABLE PROTEASE; PROMOTES ADHERENCE AND INVASION BY
CC       DIRECTLY BINDING TO A HOST CELL STRUCTURE (BY SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: Secreted (Potential).
CC   -!- DOMAIN: THE SIGNAL PEPTIDE GUIDE THE PRECURSOR TO THE PERIPLASMIC
CC       SPACE, AND THE CARBOXY-TERMINAL HELPER DOMAIN ASSOCIATES WITH THE
CC       OUTER MEMBRANE TO FORM A PORE FOR EXCRETION OF THE PROTEASE
CC       DOMAIN. THE HELPER DOMAIN IS THEN RELEASED BY AUTOPROTEOLYSIS (BY
CC       SIMILARITY).
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S6 (SERINE PROTEASE).
CC   -!- CAUTION: THIS IS A CONCEPTUAL TRANSLATION; A STOP CODON HAD TO
CC       BE SKIPPED IN POSITION 710 TO PRODUCE THIS ORF.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U32710; -; NOT_ANNOTATED_CDS.
DR   TIGR; HI0248; -.
DR   InterPro; IPR000710; IgA_S6.
DR   Pfam; PF02395; IGA1; 1.
DR   PRINTS; PR00921; IGASERPTASE.
KW   Hydrolase; Serine protease; Transmembrane; Zymogen; Signal;
KW   Complete proteome.
FT   SIGNAL        1     25       POTENTIAL.
FT   CHAIN        26      ?       ADHESION AND PENETRATION PROTEIN.
FT   PROPEP        ?   1409       HELPER PEPTIDE (POTENTIAL).
FT   ACT_SITE    250    250       BY SIMILARITY.
SQ   SEQUENCE   1409 AA;  156797 MW;  63ABC893FA84D16E CRC64;


alignment_scores:
           Quality:  121.00            Length:    650
             Ratio:  0.448              Gaps:     33
 Percent Similarity:  41.538   Percent Identity: 19.692

alignment_block:
 US-09-303-518D-465 x HAP1_HAEIN   ..

 Align seg 1/1  to: HAP1_HAEIN  from: 1  to: 1409

      94 TTTATCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCC..... 132
         |||      |||::::::   |||:::::::::::::
     119 PheThrTyrGlnIleValLysArgAsnAsnTyrGlnAlaTrpGluArgLy 135

     133 .....GACGGGAAATACCACCTA..... 150
                 ||||||   |||||||:::
     135 sHisProTyrAspGlyAspTyrHisMetProArgLeuHisLysPheValT 152

     151 ..............................................TTC 153
                                                      :::
     152 hrGluAlaGluProValGlyMetThrThrAsnMetAspGlyLysValTyr 168

     154 GGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGA.. 201
         ::::::|||   :::      ||||||      :::||||||   |||
     169 AlaAspArgGluAsnTyrProGluArgVal...ArgIleGlySerGlyAr 184

     202 ................................AACATACAAAGCCAT. 216
                                         |||:::::::|||
     184 gGlnTyrTrpArgThrAspLysAspGluGluThrAsnValHisSerSerT 201

     217 ............................CAGTTGGGCAACCTGTTCATC 237
                                      ||||||

     201 yrTyrValSerGlyAlaTyrArgTyrLeuThrAlaGlyAsnThrHisThr 217

     238 CAGCAGGCGGCCATTAAAGGAAATATC.........GGCTACATTGTCCG 278
         |||::::::      :::|||::::::         |||   :::|||
     218 GlnSerGlyAsnGlyAsnGlyThrValAsnLeuSerGlyAsnValValSe 234

     279 CTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCT 328
             :::   :::|||               |||:::   :::   :::|
     234 rProAsnHisTyrGly..............ProLeuProThrGlyGlyS 246

     329 CACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTT 378
         ||   :::||||||         |||||||||:::      :::::
     246 erLysGlyAspSer.........GlySerProMetPheIleTyrAspAla 259

     379 TACCGCATCCATTGG.....................GACGGATACGAACA 407
            :::   :::|||                     |||
     260 LysLysLysGlnTrpLeuIleAsnAlaValLeuGlnThrGly..... 273

     408 CCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTC 457
          ||||||      |||              :::|||:::|||:::
     274 .HisProPhePheGly.............ArgGlyAsnGlyPheGlnLeuI 286

     458 CCAAAGGCGCG.........AGGGATATATACAGCTACGACATAAAAGGC 498
             :::              :::::::::   :::   |||      :::
     286 leArgGluGluTrpPheTyrAsnGluValLeuAlaValAspThrProSer 302

     499 GTTGCCCAAAATATC......CGCCTCAAC..............CTGAC 527
         |||   |||:::           :::|||                :::::
     303 ValPheGlnArgTyrIleProProIleAsnGlyHisTyrSerPheValSe 319

     528 CGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCG 577
         ::::|||   :::||||||:::      :::   |||      :::   |
     319 rAsnAsnAspGlyThrGlyLysLeuThrLeuThrArgProSerLysAspG 336

     578 GTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGA 627
         |||||         :::   ||||||          :::
     336 lySerLysAlaLysSerGluValGly............ThrValLysLeu 348

     628 TACAGCCCCGAGCTGGACAGATCGGGC............AATGCCGCCGA 665
         ::::::|||:::|||:::::::::::::           :::||||||
     349 PheAsnProSerLeuAsnGlnThrAlaLysGluHisValLysAlaAlaAl 365

     666 AGCTTTCAAC..................................... 675
          ::::::|||
     365 aGlyTyrAsnIleTyrGlnProArgMetGluTyrGlyLysAsnIleTyrL 382

     676 .................GGCACTGCAGATATCGTCAAAAACATCATCGGC 708
                         ||||||      |||   :::||||||
     382 euGlyAspGlnGlyLysGlyThrLeuThrIleGluAsnAsnIleAsnGln 398

     709 GCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGA 758
         :::||||||   :::          |||:::   |||      ::::::
     399 GlyAlaGlyGlyLeuTyrPheGluGlyAsnPheVal......ValLysGl 413

     759 AGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAA 808
         :::|||      :::   :::|||   |||:::          ::::
     413 yLysGlnAsnAsnIleThrTrpGlnGlyAlaGlyValSerIleGlyGlnA 430

     809 ACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTCAAAGAC 858
         ::
     430 sp.......................................... 430

     859 TATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGC 908
                 |||      :::|||   |||::||||||:::
     431 ............AlaThrValGluTrpLysValHisAsnProGluAsn.. 442

     909 ACAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCA 958
                 :::   :::|||:::|||      :::::::::   |||:
     443 ..........AspArgLeuSerLysIleGlyIleGlyThrLeuLeuValA 456

     959 AAGGGATTGGA.....................GCTGTTCGGGGAAAATAC 987
         ::|||   |||                     |||   :::||||||
     456 snGlyLysGlyLysAsnLeuGlySerLeuSerAlaGlyAsnGlyLysVal 472
```

```
 988 GGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGA 1037
     |||         :::   :::      :::::::::   :::   ||
 473 IleLeuAspGlnGlnAlaAspGluAlaGlyGlnLysGlnAlaPheLysGl 489

1038 GATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATG 1087
     |:::::::::   :::|||:::::::::::|||:::   |||   :::|||
 489 uValGlyIleValSerGlyArgAlaThrValGlnLeuAsnSerThrAspG 506

1088 CGGCA..............TACGCCAAATACCCGTCCCCTTACCATTCC 1122
       :::               |||      :::   :::         :::
 506 lnValAspProAsnAsnIleTyrPheGlyPheArgGlyGlyArgLeuAsp 522

1123 CGAAATATCCGTTCAAACTTGGAGCAGCGTTAC..................... 1155
       |||   :::|||         :::|||
 523 LeuAsnGlyHisSerLeuThrPheLysArgIleGlnAsnThrAspGluGl 539

1155 ..................................................... 1155

 539 yAlaMetIleValAsnHisAsnThrThrGlnValAlaAsnIleThrIleT 556

1156 ..GGCAAAGAAAACATCACCTCCTCAACCGTGCCGCCGTCAAACGGAAAG 1203
       |||:::|||:::||||||:::            |||||||||   |||
 556 hrGlyAsnGluSerIleThrAla...........ProSerAsnLysLys 568

1204 AATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTGA 1253
     |||:::         ||||||         :::   ::::::   :::::
 569 AsnIle.........AsnLysLeuAspTyrSerLysGluIleAlaTyrAs 582

1254 CGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAA 1303
     :|||      |||   :::   :::|||:::   |||::::::   |||:
 582 nGly...TrpPheGlyGluThrAspLysAsn...LysHisAsnGlyArgL 597

1304 TTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGTCTTTAAT 1353
     ::|||                        |||   :::|||:::   :::
 597 euAsnLeu.....................IleTyrLysProThrThrGlu 606

1354 CCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAAT 1403
       :::      :::   |||:::   :::
 607 AspArgThrLeuLeuLeuSerGlyGlyThr..................... 616

1404 TCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCTA 1453
                          :::|||   |||                |
 617 ..............AsnLeuLysGlyAspIleThrGlnThrLysGlyL 628

1454 AAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAATGGATAT 1503
     ||   :::      :::   |||   |||   :::   |||      :::
 628 ysLeuPhePheSerGlyArgProThrProHisAlaTyrAsn......His 642

1504 TTGGATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAGAACTAAAGG 1553
     |||||||||         :::|||:::::      :::      :::||
 643 LeuAspLys.........ArgTrpSerGluMetGluGlyIleProGlnGl 656

1554 TCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACAGGAAGAGAGCAAC 1603
     |   |||:::   ||||||   :::
 656 y...GluIleValTrpAspTyrAsp........................ 663

1604 TTGGATGGGCTAGTAGGGATGGTAAGCATTTAAATATATCAATTGATGGA 1653
           |||   :::|||      |||      |||:::::|||   |||
 664 .....TrpIleAsnArgThrPheLysAlaGluAsnPheGlnIleLysGly 678

seq_name: SwissProt_40:YB23_HUMAN

seq_documentation_block:
ID   YB23_HUMAN     STANDARD;      PRT;   768 AA.
AC   Q9ULJ7;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Hypothetical protein KIAA1223 (Fragment).
GN   KIAA1223.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=20039619; PubMed=10574462;
RA   Nagase T., Ishikawa K.-I., Kikuno R., Hirosawa M., Nomura N.,
RA   Ohara O.;
RT   "Prediction of the coding sequences of unidentified human genes. XV.
RT   The complete sequences of 100 new cDNA clones from brain which code
RT   for large proteins in vitro.";
RL   DNA Res. 6:337-345(1999).
CC   -!- SIMILARITY: CONTAINS AT LEAST 14 ANK REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AB033049; BAA86537.1; -.
DR   HSSP; P42771; 1BI7.
DR   InterPro; IPR002110; ANK.
DR   Pfam; PF00023; ank; 13.
DR   SMART; SM00248; ANK; 13.
DR   PROSITE; PS50088; ANK_REPEAT; 13.
DR   PROSITE; PS50297; ANK_REP_REGION; 1.
KW   Hypothetical protein; Repeat; ANK repeat.
FT   NON_TER       1      1
FT   REPEAT        1     11       ANK 1.
FT   REPEAT       15     44       ANK 2.
FT   REPEAT       48     82       ANK 3.
FT   REPEAT       86    115       ANK 4.
FT   REPEAT      119    148       ANK 5.
FT   REPEAT      152    181       ANK 6.
FT   REPEAT      185    214       ANK 7.
FT   REPEAT      218    247       ANK 8.
FT   REPEAT      251    280       ANK 9.
FT   REPEAT      284    313       ANK 10.
FT   REPEAT      317    346       ANK 11.
FT   REPEAT      350    379       ANK 12.
FT   REPEAT      383    412       ANK 13.
FT   REPEAT      416    446       ANK 14.
SQ   SEQUENCE   768 AA;  82819 MW;  2913B69BE2DFE06D CRC64;


alignment_scores:
           Quality:  120.50              Length:    668
             Ratio:  0.429                 Gaps:     33
 Percent Similarity:  42.066   Percent Identity: 20.659

alignment_block:
 US-09-303-518D-465 x YB23_HUMAN   ..

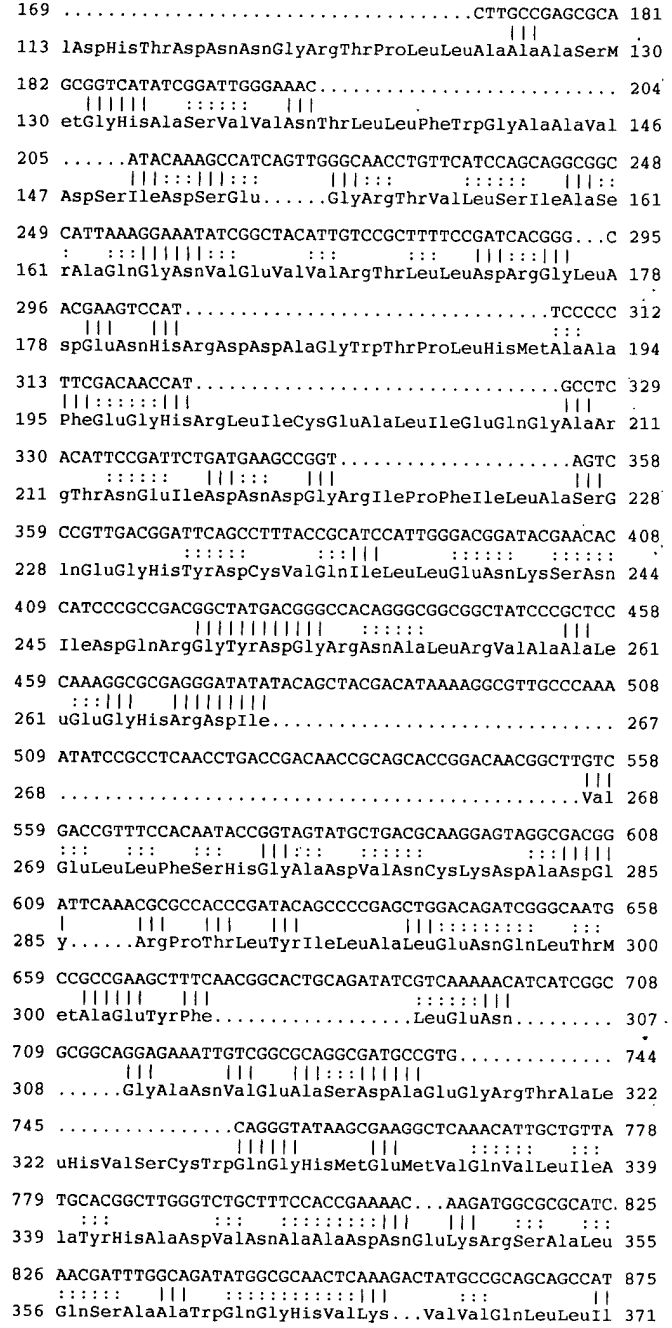
 Align seg 1/1  to: YB23_HUMAN  from: 1  to: 768

  31 CTGTCCATACTGGCAGTGTGCCTGCCG...ATGCATGCACACGCCTCAGA 77
     ||||||:::   |||:::|||:::|||         :::||||||
  53 LeuSerValAlaAlaLeuCysValProAlaSerLysGlyHisAla..... 67

  78 TTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGACCGT.......... 117
                 |||   :::      ::::::||||||
  68 ............SerValValSerLeuLeuIleAspArgGlyAlaGluV 80

 118 ..CAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGGG 165
       :::|||   :::   ||||||         |||:::   :::
  80 alAspHisCysAspLysAspGlyMetThrProLeuLeuValAlaAlaTyr 96

 166 GAA...................................................... 168
     |||
  97 GluGlyHisValAspValValAspLeuLeuLeuGluGlyGlyAlaAspVa 113
```


169 CTTGCCGAGCGCA 181
113 lAspHisThrAspAsnAsnGlyArgThrProLeuLeuAlaAlaAlaSerM 130
182 GCGGTCATATCGGATTGGGAAAC..... 204
130 etGlyHisAlaSerValValAsnThrLeuLeuPheTrpGlyAlaAlaVal 146
205 ATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGC 248
147 AspSerIleAspSerGlu......GlyArgThrValLeuSerIleAlaSe 161
249 CATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGG...C 295
161 rAlaGlnGlyAsnValGluValValArgThrLeuLeuAspArgGlyLeuA 178
296 ACGAAGTCCAT.....TCCCCC 312
178 spGluAsnHisArgAspAspAlaGlyTrpThrProLeuHisMetAlaAla 194
313 TTCGACAACCAT.....GCCTC 329
195 PheGluGlyHisArgLeuIleCysGluAlaLeuIleGluGlnGlyAlaAr 211
330 ACATTCCGATTCTGATGAAGCCGGT.....AGTC 358
211 gThrAsnGluIleAspAsnAspGlyArgIleProPheIleLeuAlaSerG 228
359 CCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACAC 408
228 lnGluGlyHisTyrAspCysValGlnIleLeuLeuGluAsnLysSerAsn 244
409 CATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCC 458
245 IleAspGlnArgGlyTyrAspGlyArgAsnAlaLeuArgValAlaAlaLe 261
459 CAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAA 508
261 uGluGlyHisArgAspIle..... 267
509 ATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTC 558
268 Val 268
559 GACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGG 608
269 GluLeuLeuPheSerHisGlyAlaAspValAsnCysLysAspAlaAspGl 285
609 ATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATG 658
285 y......ArgProThrLeuTyrIleLeuAlaLeuGluAsnGlnLeuThrM 300
659 CCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGC 708
300 etAlaGluTyrPhe.....LeuGluAsn..... 307
709 GCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTG..... 744
308 GlyAlaAsnValGluAlaSerAspAlaGluGlyArgThrAlaLe 322
745 CAGGGTATAAGCGAAGGCTCAAACATTGCTGTTA 778
322 uHisValSerCysTrpGlnGlyHisMetGluMetValGlnValLeuIleA 339
779 TGCACGGCTTGGGTCTGCTTTCCACCGAAAAC...AAGATGGCGCGCATC. 825
339 laTyrHisAlaAspValAsnAlaAlaAspAsnGluLysArgSerAlaLeu 355
826 AACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCAT 875
356 GlnSerAlaAlaTrpGlnGlyHisValLys...ValValGlnLeuLeuIl 371
876 CCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCG 925
371 eGluHisGlyAlaValValAspHisThrCysAsnGlnGlyAlaThrAlaL 388
926 TCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTT 975
388 euCys.....IleAlaAlaGln 393
976 CGGGGAAAATACGGCTTGGGCGGCATCACG.....GCACATCC 1013
394 GluGlyHisIleAspValValGlnValLeuLeuGluHisGlyAlaAspPr 410
1014 TGTCAAGCGGTCGCAGATGGGC.....GAGATCGCATTGCCGA 1051
410 oAsnHisAlaAspGlnPheGlyArgThrAlaMetArgValAlaAlaLysA 427
1052 AAGGGAAATCCGCCGTC.....AGCGACAATTTTGCCGATGCG 1089
427 snGlyHisSerGlnIleIleLysLeuLeuGluLysTyrGlyAlaSerSer 443
1090 GCATACGCCAAATACCCGTCCCCTTACCATTCC..... 1122
444 LeuAsnGlyCysSerProSerProValHisThrMetGluGlnLysProLe 460
1123 .CGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCA 1171
460 uGlnSerLeuSerSerLysValGlnSerLeuThrIleLysSerAsnSerS 477
1172 CCTCCTCAACC.....GTGCCGCCGTCA..... 1194
477 erGlySerThrGlyGlyGlyAspMetGlnProSerLeuArgGlyLeuPro 493
1195 AACGGAAAGAATGTGAAACTGGCAAAC...AAACGCCACCCGAAGACCAA 1241
494 AsnGlyProThrHisAlaPheSerSerProSerGluSerProAspSerTh 510
1242 AGTGCCGTTTGACGGTAAAGGGTTTCCGAAT.....TTTGAAA 1279
510 rValAspArgGlnLysSerSerLeuSerAsnAsnSerLeuLysSerSerL 527
1280 AAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAAT 1329
527 ysAsnSerSerLeuArgThrThrSerSerThrAlaThrAlaGlnThrVal 543
1330 CCTATAGATGAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCA 1379
544 ProIleAsp.....SerPheHi 549
1380 TTCTTGGTCTATAACTGCCAGAATTCAATACGCAAAATTACCAAGG...C 1426
549 sAsnLeuSerPheThrGluGlnIleGlnGlnHisSerLeuProArgSerA 566
1427 AAGGTAGAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCG 1476
566 rgSerArgGlnSerIleValSerProSerSerThrThrGlnSer..... 580
1477 CTACCAAAAGGACCTAATAATGGATATTTGGATAAATTTGGTAATGAATG 1526
581 LeuGlyGlnSerHisAsnSer..... 587
1527 GACTAAAGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATGG..... 1569
588 ProSer.....SerGluPheGluTrpSerGlnV 597
1570 GATGTTCAATTGTCTAAAACAGGAAGAGAGCAA 1602
597 alLysProSerLeuLysSerThrLysAlaSerLysGlyGlyLysSerGlu 613
1603 CTTGGATGGGCTAGTAGGGATGGTAAGCATTTAAATAT 1640
614 AsnSerAlaLysSerGlySerAlaGlyLysLysAlaLysGlnSerAsnSe 630
1641 ATCA 1644

```
        |||
    630 rSer 631

seq_name: SwissProt_40:P115_CHICK

seq_documentation_block:
ID   P115_CHICK     STANDARD;      PRT;   762 AA.
AC   Q98917;
DT   15-JUL-1998 (Rel. 36, Created)
DT   15-JUL-1998 (Rel. 36, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   115 kDa melanosomal matrix protein precursor.
GN   MMP115.
OS   Gallus gallus (Chicken).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Archosauria; Aves; Neognathae; Galliformes; Phasianidae; Phasianinae;
OC   Gallus.
OX   NCBI_TaxID=9031;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=WHITE LEGHORN; TISSUE=Retinal pigment epithelium;
RX   MEDLINE=92020667; PubMed=1924173;
RA   Mochii M., Agata K., Eguchi G.;
RT   "Complete sequence and expression of a cDNA encoding a chicken
RT   115-kDa melanosomal matrix protein.";
RL   Pigment Cell Res. 4:41-47(1991).
RN   [2]
RP   CHARACTERIZATION.
RC   STRAIN=WHITE LEGHORN; TISSUE=Retinal pigment epithelium;
RX   MEDLINE=88311098; PubMed=3409326;
RA   Mochii M., Agata K., Kobayashi H., Yamamoto T.S., Eguchi G.;
RT   "Expression of gene coding for a melanosomal matrix protein
RT   transcriptionally regulated in the transdifferentiation of chick
RT   embryo pigmented epithelial cells.";
RL   Cell Differ. 24:67-74(1988).
CC   -!- FUNCTION: MIGHT BE REQUIRED FOR POLYMERIZATION OF MELANIN ONTO THE
CC       CORE STRUCTURE OF MELANOSOMES WITH ENZYMIC FUNCTION OF TYROSINASE.
CC   -!- SUBCELLULAR LOCATION: ON THE FIBROUS MATRIX STRUCTURE OF THE
CC       PREMELANOSOME.
CC   -!- TISSUE SPECIFICITY: SPECIFIC TO PIGMENTED EPITHELIAL CELLS AND
CC       MELANOCYTES. NOT EXPRESSED IN LENS, NEURAL RETINA, BRAIN, HEART,
CC       GIZZARD OR LIVER.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSEDED DURING THE REDIFFERENTIATION OF
CC       PIGMENTED EPITHELIAL CELLS (PEC).
CC   -!- PTM: GLYCOSYLATED.
CC   -!- SIMILARITY: BELONGS TO THE PMEL-17/NMB FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 PKD DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D88348; BAA13589.1; -.
DR   InterPro; IPR000601; PKD_domain.
DR   Pfam; PF00801; PKD; 1.
DR   SMART; SM00089; PKD; 1.
DR   PROSITE; PS50093; PKD; 1.
KW   Signal; Glycoprotein; Repeat.
FT   SIGNAL        1     19       POTENTIAL.
FT   CHAIN        20    762       115 KDA MELANOSOMAL MATRIX PROTEIN.
FT   DOMAIN      223    323       PKD.
FT   DOMAIN      441    532       4 X 20-24 AA APPROXIMATE TANDEM REPEATS.
FT   REPEAT      441    464       1.
FT   REPEAT      465    488       2.
FT   REPEAT      489    508       3.
FT   REPEAT      509    532       4.
FT   CARBOHYD    111    111       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    115    115       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    346    346       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    651    651       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    659    659       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   762 AA;  77356 MW;  172C8DDB4FDCE7C6 CRC64;


alignment_scores:
            Quality:  120.00              Length:    571
              Ratio:  0.498                Gaps:     29
 Percent Similarity:  42.207   Percent Identity: 21.366

alignment_block:
 US-09-303-518D-465 x P115_CHICK   ..

 Align seg 1/1  to: P115_CHICK  from: 1  to: 762

    113 ACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGG 162
        |||:::          :::||||||   :::   ||||||      |||
    274 ThrLeuIleSerArgSerProThrValThrHisThrTyrLeuGlnAla.. 289

    163 GGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAG 212
         |||:::::::   :::      |||::::::                 |
    290 .GlySerPheAlaAlaArgLeuValLeuGln.............AlaA 301

    213 CCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATA 262
        |||||      :::::::|||:::::|||      |||:::   :::
    301 laIleProLeuSerSerCysGlyThrSerAlaProProValValAsp... 316

    263 TCGGCTACATTGTCCGCTTTTCCGATCACGGGC...ACGAAGTCCATTCC 309
                           |||   ||||||   :::   |||:::
    317 ....................ProThrThrGlyProValProSerLeuGl 326

    310 CCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCC 359
         |||:::::::|||                        :::||||||   |
    326 yProThrAlaThr.......................GlnProValGlyP 335

    360 CGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACC 409
        ||      :::::::::   ||||||
    335 roThrGlySerGlyThrAlaThrAla........................ 343

    410 ATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCG..... 454
              |||:::   :::||||||      :::||||||   |||
    344 ......ProSerAsnLeuThrGlySerGlyThrAlaAlaAlaProGlyTh 358

    455 .......CTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGG 497
                  |||:::|||   |||      :::      |||      :
    358 rThrAlaAlaProArgAlaSerGlyAlaProAlaGluProThrGlyValS 375

    498 CGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGAC 547
        :::::      :::|||   |||         |||          ||||||
    375 erValAlaValLeuSerAspSerMet...GluThrGluProLeuProAsp 390

    548 AACGGCTTGTCG.................................AC 561
              ||||||                                  ||
    391 ProValLeuSerThrAlaValAlaAspAlaAlaAlaGlyThrAspProTh 407

    562 CGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATT 611
        |::::::   :::|||   :::            :::   ::::::   :
    407 rAlaAspProLeuProProThr...........SerValSerSerGlyG 420

    612 CAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCG 661
        ::::::||||||      :::|||||:::   ::::::   ::::::
    420 lyAspAlaProGlyThrValAlaProThrAlaValGluGlySerValAla 436

    662 CCGAAGCTTTCAACGGCACT............................ 681
              ::::::||||||
    437 AlaGlyValGlyThrAlaGluAspValAlaAlaAlaThrProGlyAlaTh 453

    682 ....GCAGATATCGTCAAAAACATCATCGGCGCGGCA...GGAGAAATTG 724
            ||||||::::::   :::      |||||||::   |||:::   |
    453 rAlaAlaAspValAlaValAspThrAlaGlyAlaThrAspGlyAspAlaV 470

    725 TCGGC.........................GCAGGCGATGCCGTGCAG 747
        |||||                          ::::|||   :::
```

```
 470 alGlyProThrAlaAlaAlaThrAlaGluSerIleAlaAspProThrAla 486

 748 GGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCT 797
     |||   ::::::|||:::   ::::::
 487 GlyAlaThrAspGlyAspAlaValGly..................... 495

 798 TTCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGC 847
                     |||      :::::::::||||||
 496 ..................AlaThrAlaGluSerIleAlaAsp..... 503

 848 AACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAAC 897
                    :::|||:::      |||      :::   :::
 504 .............ProThrAlaGlyAlaThrAspGlyAspAlaValGly 515

 898 CCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATCTTTACGGCAGT 947
     |||:::||||||        |||:::::::::::::
 516 ProThrAlaAlaAlaThrAlaGluSerIleAlaAsp.............. 527

 948 CATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCG 997
         |||:::   |||      ||||||   :::      :::   :::|
 528 ....ProThrAlaGlyAlaThrAlaValSerSerGlySerAlaThrAlaG 543

 998 GCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTG 1047
     ||   ||||||:::|||                              :::   |||
 543 lyAlaThrAlaGluPro.....................LeuLeuLeu 551

1048 CCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGC 1097
        |||   ::::::      ::::::      :::   :::   |||
 552 ValLysArgGlnAlaProGluAlaGluProThrGlyCysValLeuTyrAr 568

1098 CAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGC 1147
                     |||   :::   :::   :::   |||:::   |
 568 g.............TyrGlyThrPheSerThrGluLeuAsnIleValG 580

1148 AGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTG............ 1185
     ||      |||   |||::::::::::         |||
 580 ln......GlyIleGluSerValAlaIleValGlnValValProAlaAla 594

1186 CCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAA 1235
     |||   ::::::|||::::::|||:::|||:::
 595 ProGluGlySerGlyAsnSerValGluLeuThr.................. 605

1236 GACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACG 1285
             |||      :::|||   ::::::|||:::           |
 606 .......ValThrCysGluGly...SerLeuProGluGluValCysThrV 619

1286 TAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATA 1335
     ||      |||::::::   :::||||||      |||:::
 619 alValAlaAspAlaGluCysArgThrAlaGlnMetGlnThrCysSerAla 635

1336 GATGAACCCGTC............................TTTAATCC 1355
           |||:::                            ||||||
 636 ValAlaProAlaProGlyCysGlnLeuValLeuArgGlnAspPheAsnGl 652

1356 TAAAGGT.........................................T 1363
      :::|||                                        :
 652 nSerGlyLeuTyrCysLeuAsnValSerLeuAlaAsnGlyAsnGlyLeuA 669

1364 CTGTCGGATCGGCTCAT......TCTTGGTCTATAACTGCCAGAATTCAA 1407
     ::|||:::|||:::|||              ||||||   :::|||   |||
 669 laValAlaSerThrHisValAlaValGlySerIleProSerArg...Gln 684

1408 TACGCAAAATTACCAAGGCAAGGT........................... 1431
     :::             |||   :::
 685 TrpHisHisAlaHisArgGlyAlaAlaLeuGlyThrAlaHisGlyArgCy 701

1432 ..........AGAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAG 1471
               |||      :::::||||||   :::   :::|||:::|
 701 sSerGlyHisArgCysLeuHisLeuProProCysGluValGlnProAlaA 718

1472 CACCG...CTACCAAAAGGACCTAATAATGGATATTTGGATAAA...... 1512
     ||         |||   ||||||            |||
 718 laAlaHisSerProHisGlyProProAlaProGlnLeuAlaAlaProArg 734

     1513 ..................................TTTGGTAATGAATG 1526
                                            |||::::::||
      735 CysTyrProAlaPheAlaAlaAlaProGlyPheTrpGlyGlySerGlnTr 751

     1527 GACTAAAGGTCCA 1539
          |   |||   |||
      751 pArgLysGlnPro 755
```

```
seq_name: SwissProt_40:SON_MOUSE

seq_documentation_block:
ID   SON_MOUSE      STANDARD;      PRT;  2404 AA.
AC   Q9QX47; Q9QXP5; Q9CQK6; Q9CQ12;
DT   01-MAR-2002 (Rel. 41, Created)
DT   01-MAR-2002 (Rel. 41, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   SON protein.
GN   SON.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A. (ISOFORMS 1 AND 2).
RC   STRAIN=129/Sv;
RX   MEDLINE=20408886; PubMed=10950926;
RA   Wynn S.L., Fisher R.A., Pagel C., Price M., Liu Q.Y., Khan I.M.,
RA   Zammit P., Dadrah K., Mazrani W., Kessling A., Lee J.S., Buluwela L.;
RT   "Organization and conservation of the GART/SON/DONSON locus in mouse
RT   and human genomes.";
RL   Genomics 68:57-62(2000).
RN   [2]
RP   SEQUENCE OF 1-116 FROM N.A.
RC   STRAIN=C57BL/6J; TISSUE=Hippocampus, Small intestine, and Tongue;
RX   MEDLINE=21085660; PubMed=11217851;
RA   Kawai J., Shinagawa A., Shibata K., Yoshino M., Itoh M., Ishii Y.,
RA   Arakawa T., Hara A., Fukunishi Y., Konno H., Adachi J., Fukuda S.,
RA   Aizawa K., Izawa M., Nishi K., Kiyosawa H., Kondo S., Yamanaka I.,
RA   Saito T., Okazaki Y., Gojobori T., Bono H., Kasukawa T., Saito R.,
RA   Kadota K., Matsuda H.A., Ashburner M., Batalov S., Casavant T.,
RA   Fleischmann W., Gaasterland T., Gissi C., King B., Kochiwa H.,
RA   Kuehl P., Lewis S., Matsuo Y., Nikaido I., Pesole G., Quackenbush J.,
RA   Schriml L.M., Staubli F., Suzuki R., Tomita M., Wagner L., Washio T.,
RA   Sakai K., Okido T., Furuno M., Aono H., Baldarelli R., Barsh G.,
RA   Blake J., Boffelli D., Bojunga N., Carninci P., de Bonaldo M.F.,
RA   Brownstein M.J., Bult C., Fletcher C., Fujita M., Gariboldi M.,
RA   Gustincich S., Hill D., Hofmann M., Hume D.A., Kamiya M., Lee N.H.,
RA   Lyons P., Marchionni L., Mashima J., Mazzarelli J., Mombaerts P.,
RA   Nordone P., Ring B., Ringwald M., Rodriguez I., Sakamoto N.,
RA   Sasaki H., Sato K., Schoenbach C., Seya T., Shibata Y., Storch K.-F.,
RA   Suzuki H., Toyo-oka K., Wang K.H., Weitz C., Whittaker C., Wilming L.,
RA   Wynshaw-Boris A., Yoshida K., Hasegawa Y., Kawaji H., Kohtsuki S.,
RA   Hayashizaki Y.;
RT   "Functional annotation of a full-length mouse cDNA collection.";
RL   Nature 409:685-690(2001).
CC   -!- FUNCTION: Transcriptional repressor. Binds to the consensus DNA
CC       sequence: 5'-GA[GT]AN[CG][AG]CC-3'. Might protect cells from
CC       apoptosis. Might be involved in pre-mRNA splicing (By similarity).
CC   -!- SUBCELLULAR LOCATION: Nuclear (By similarity).
CC   -!- ALTERNATIVE PRODUCTS: 2 isoforms; 1 (shown here) and 2; are
CC       produced by alternative splicing.
CC   -!- TISSUE SPECIFICITY: Widely expressed.
CC   -!- DOMAIN: Contains 8 types of repeats which are distributed in 3
CC       regions.
CC   -!- SIMILARITY: CONTAINS 1 G-PATCH DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 DRBM (DOUBLE-STRANDED RNA-BINDING) DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
```

```
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ------------------------------------------------------------------------
DR   EMBL; AF193606; AAF23120.1; -.
DR   EMBL; AF193595; AAF23120.1; JOINED.
DR   EMBL; AF193596; AAF23120.1; JOINED.
DR   EMBL; AF193597; AAF23120.1; JOINED.
DR   EMBL; AF193598; AAF23120.1; JOINED.
DR   EMBL; AF193599; AAF23120.1; JOINED.
DR   EMBL; AF193600; AAF23120.1; JOINED.
DR   EMBL; AF193601; AAF23120.1; JOINED.
DR   EMBL; AF193602; AAF23120.1; JOINED.
DR   EMBL; AF193603; AAF23120.1; JOINED.
DR   EMBL; AF193604; AAF23120.1; JOINED.
DR   EMBL; AF193605; AAF23120.1; JOINED.
DR   EMBL; AF193607; AAF23121.1; -.
DR   EMBL; AK019312; BAB31659.1; -.
DR   EMBL; AK019081; BAB31536.1; -.
DR   EMBL; AK008478; BAB25691.1; -.
DR   EMBL; AK008256; BAB25562.1; -.
DR   MGD; MGI:98353; Son.
DR   InterPro; IPR001159; DS_RBD.
DR   InterPro; IPR000467; G_patch.
DR   Pfam; PF00035; dsrm; 1.
DR   Pfam; PF01585; G-patch; 1.
DR   SMART; SM00443; G-patch; 1.
DR   PROSITE; PS50137; DS_RBD; 1.
DR   PROSITE; PS50174; G_PATCH; 1.
KW   RNA-binding; DNA-binding; Nuclear protein; Repeat;
KW   Alternative splicing.
FT   DOMAIN      721    850       13 X 10 AA TANDEM REPEATS OF L-A-[ST]-
FT                                [NSG]-[TS]-MDSQM.
FT   DOMAIN      867    943       11 X 7 AA TANDEM REPEATS OF [DR]-P-Y-R-
FT                                [LI][AG][QHP].
FT   DOMAIN      961   1080       14 X 6 AA REPEATS OF [ED]-R-S-M-M-S.
FT   REPEAT      961    966       1-1.
FT   REPEAT      969    974       1-2.
FT   REPEAT      976    981       1-3.
FT   REPEAT      985    990       1-4.
FT   REPEAT      993    998       1-5.
FT   REPEAT     1001   1006       1-6.
FT   REPEAT     1010   1015       1-7.
FT   REPEAT     1018   1023       1-8.
FT   REPEAT     1026   1031       1-9.
FT   REPEAT     1035   1040       1-10.
FT   REPEAT     1044   1049       1-11.
FT   REPEAT     1055   1060       1-12.
FT   REPEAT     1066   1071       1-13.
FT   REPEAT     1075   1080       1-14.
FT   DOMAIN     1101   1133       3 X 11 AA TANDEM REPATS OF P-P-L-P-P-E-E-
FT                                P-P-[TME]-[MTG].
FT   DOMAIN     1910   1979       7 X 7 AA REPEATS OF P-S-R-R-S-R-[TS].
FT   REPEAT     1910   1916       2-1.
FT   REPEAT     1938   1944       2-2.
FT   REPEAT     1945   1951       2-3.
FT   REPEAT     1952   1958       2-4.
FT   REPEAT     1959   1965       2-5.
FT   REPEAT     1966   1972       2-6.
FT   REPEAT     1973   1979       2-7 (APPROXIMATE).
FT   DOMAIN     1919   1990       2 X 19 AA REPEATS OF P-S-R-R-R-R-S-R-S-V-
FT                                V-R-R-R-S-F-S-I-S.
FT   REPEAT     1919   1937       3-1.
FT   REPEAT     1980   1990       3-2 (APPROXIMATE).
FT   DOMAIN     1991   2017       3 X TANDEM REPEATS OF [ST]-P-[VLI]-R-
FT                                [RL]-[RK]-[RF]-S-R.
FT   DOMAIN     2283   2329       G-PATCH.
FT   DOMAIN     2349   2404       DRBM.
FT   VARSPLIC   2086   2086       K -> F (IN ISOFORM 2).
FT   VARSPLIC   2087   2404       MISSING (IN ISOFORM 2).
SQ   SEQUENCE   2404 AA;  261428 MW;  648BF28ED3FC01D9 CRC64;

alignment_scores:
```

```
          Quality:  120.00          Length:    585
            Ratio:  0.504             Gaps:     28
Percent Similarity:  40.684   Percent Identity: 21.197

alignment_block:
 US-09-303-518D-465 x SON_MOUSE   ..

 Align seg 1/1  to: SON_MOUSE  from: 1  to: 2404

     102 GCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTAT 151
         |||   :::|||   |||      :::::::::|||:::
    1832 AlaArgLysArgSerSerLysSerLysSerHisArgSer..... 1844

     152 TCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGA 201
           |||      :::      |||:::::::|||      |||   ::::
    1845 ....GlnThrArgSerArgSerArgSerArgArgArgArgArgSerSerA 1860

     202 AACATACAAAGCCATCAGTTGGGCAA..................... 227
         ::      :::   |||   |||:::
    1860 rgSerArgSerLysSerArgGlyArgArgSerValSerLysGluLysArg 1876

     228 .........CCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCT 268
                  |||   |||   :::   ::::::   ||||||   :::
    1877 LysArgSerProLysHisArgSerLysSerArgGluArgLysArgLys.. 1892

     269 ACATTGTCCGCTTTTCCGATCA.......................C 291
                          ||||||                      |
    1893 ..............ArgSerSerSerArgAspAsnArgLysAlaAlaA 1904

     292 GGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTC 341
         ||||||||||      |||   |||:::
    1904 rgAlaArgSerArgThrProSerArgArg..................... 1913

     342 TGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATT 391

    1913 ..................................................... 1913

     392 GGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGC 441
         :::|||   :::||||||||||||||||||                   |
    1914 .SerArgSerHisThrProSerArgArgArg.................A 1924

     442 GGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACAT 491
         ||   :::|||   ::::::||||||:::      |||:::      :::
    1924 rgSerIleSerValGlyArgArgArgSerPheSerIleSerProSerArg 1940

     492 AAAAGGCGTTGCCCA.............AAATATCCGCCTCAACCTGACCG 529
         :::   |||   |||             :::   |||   :::   :::||
    1941 ArgSerArgThrProSerArgArgSerArgThrProSerArgArgSerAr 1957

     530 ACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGT 579
         |   |||:::::::|||:::   :::           |||:::   |||
    1957 gThrProSerArgArgSerArgThr.............ProSerArgArgS 1970

     580 AGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATA 629
               :::   :::|||   ||||||          :::|||      :::
    1970 erArgThrProSerArgArgArgArg.........SerArgSerAlaVal 1983

     630 CAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCA 679
         :::   |||:::   :::|||:::      |||
    1984 ArgArgArgSerPheSerIleSerProValArgLeu.............. 1995

     680 CTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGC 729
                                      ||||||:::|||         ||||
    1996 .......................ArgArgSerArgThrProLeuArgA 2004

     730 GCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTAT 779
         |||||      |||:::      :::|||:::
    2004 rgArgPheSerArgSerProIleArgArgLys.................. 2014

     780 GCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACG 829
            |||      |||:::   :::   ::::::       |||::::::|
    2015 ...ArgSerArgSerSerGluArgGlyArgSerPro.LysArgLeuThrA 2030
```

```
 830 ATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGC 879
     |||||  |||  |||||||||  :::  |||  |||
2030 spLeu...AspLysAlaGlnLeuLeuGluIleAlaLysAla..... 2042

 880 GATTGGGCAGTCCAAAACCCCAATGCCGCA.............CAAGG 914
                     ||||||||                    ||
2043 ..................AsnAlaAlaAlaMetCysAlaLysAlaGl 2052

 915 CATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGA 964
     |:::          |||:::     |||      |||
2052 yValProLeuProProAsnLeuLysProAlaProProPro..... 2065

 965 TTGGAGCTGTTCGGGGAAAATACGGC......TTGGGCGGCATCACGGCA 1008
       :::::::::  |||  :::         ||||||  |||
2066 .....ThrIleGluGluLysValAlaLysLysSerGlyGlyAlaThrIle 2080

1008 ................................................. 1008

2081 GluGluLeuThrGluLysCysLysGlnIleAlaGlnSerLysGluAspAs 2097

1008 ................................................. 1008

2097 pAspValIleValAsnLysProHisValSerAspGluGluGluGluGluP 2114

1009 ..............CATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCA 1044
                  ||||||  |||  |||:::        |||
2114 roProPheTyrHisHisProPheLysLeuSerGluProLysProIlePhe 2130

1045 TTG.............................CCGAAAGGGAAATC 1061
     :::                            ||||||:::::
2131 PheAsnLeuAsnIleAlaAlaAlaLysProThrProProLysSerGlnVa 2147

1062 CGCCGTCAGCGACAATTTT............................. 1080
      :::::::::   :::|||
2147 lThrLeuThrLysGluPheProValSerSerGlySerGlnHisArgLysL 2164

1081 .....GCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATTCCCGA 1125
       ||||||:::::::|||:::::::::
2164 ysGluAlaAspSerValTyrGlyGluTrpValPro.............. 2175

1126 AATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTC 1175
     :::::::::|||  |||:::          :::|||:::  ||
2176 ...ValGluLysAsnGlyGluGluSerLysAspAspAspAsnValPheSe 2191

1176 CTCAACCGTGCCGCCG..............TCAAACGGAAAGAATGTGA 1210
     ||||::::::|||                 |||::::::  :::  :
2191 rSerSerLeuProSerGluProValAspIleSerThrAlaMetSerGluA 2208

1211 AA...CTGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGT 1257
     ::  ||||||:::||||||      ::::::         :::::::::
2208 rgAlaLeuAlaGlnLysArgLeuSerGluAsnAlaPheAspLeuGluAla 2224

1258 AAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTAA 1307
       ::::::  |||                 :::  :::  ||||||::
2225 MetSerMetLeuAsn..............ArgAlaGlnGluArgIleAs 2236

1308 TACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGTCTTTAATCCTA 1357
     :  |||     |||:::|||  |||                 |||
2236 p...AlaTrpAlaGlnLeuAsnSerIle.................ProG 2246

1358 AAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTCAA 1407
           :::||||||::           :::|||  :::
2246 lyGlnPheThrGlySerThrGlyValGlnValLeuThrGlnGlu..... 2260

1408 TACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCTAAAAA 1457
       :::|||  :::  |||        :::|||          ::
2261 ......GlnLeuAlaAsnThrGlyAlaGlnAlaTrpIleLysLysAspGl 2275

1458 TTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAATGGATATTTGG 1507
     ::::     :::||||||::     |||  :::         :::
2275 nPheLeuArgAlaAlaProValThrGlyGlyMetGlyAlaValLeuMetA 2292

1508 ATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAGAACTAAAGGTCAA 1557
```

```
          |||:::|||      |||   :::|||   :::      |||::::::
2292 rgLysMetGly......TrpArgGluGlyGluGlyLeuGlyLysAsnLys 2306

1558 GAA 1560
     |||
2307 Glu 2307
```

```
seq_name: SwissProt_40:YS8A_CAEEL

seq_documentation_block:
ID   YS8A_CAEEL     STANDARD;      PRT;   796 AA.
AC   Q09625;
DT   01-NOV-1995 (Rel. 32, Created)
DT   01-NOV-1995 (Rel. 32, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Hypothetical 84.3 kDa protein ZK945.10 in chromosome II.
GN   ZK945.10.
OS   Caenorhabditis elegans.
OC   Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Rhabditoidea;
OC   Rhabditidae; Peloderinae; Caenorhabditis.
OX   NCBI_TaxID=6239;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=BRISTOL N2;
RA   Wilkinson-Sproat J.;
RL   Submitted (FEB-1995) to the EMBL/GenBank/DDBJ databases.
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; Z48544; CAA88444.1; -.
DR   EMBL; Z48582; CAA88444.1; JOINED.
DR   WormPep; ZK945.10; CE01732.
KW   Hypothetical protein; Transmembrane.
FT   TRANSMEM     11     30       POTENTIAL.
FT   DOMAIN      273    546       SER/THR-RICH.
FT   DOMAIN      656    752       SER/THR-RICH.
SQ   SEQUENCE   796 AA;  84306 MW;  76DC5B03E6357A6A CRC64;


alignment_scores:
            Quality:  119.00              Length:    428
              Ratio:  0.598                 Gaps:     16
 Percent Similarity:  46.495   Percent Identity: 21.495

alignment_block:
 US-09-303-518D-465 x YS8A_CAEEL   ..

 Align seg 1/1  to: YS8A_CAEEL  from: 1  to: 796

      68 ACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGACCGT 117
         |||      :::   :::|||      .:::            ||||||||
     287 ThrSerThrValThrSerThrThrThrValProThrSerThrSerThrVa 303

     118 C.................AGCATTTCGAACCCGACGGGAAATACCACCT 149
         |                 |||   |||:::|||:::   :::||||||
     303 lThrThrAlaMetSerThrSerThrProSerThrSerThrThrI 320

     150 ATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCG.........GTCATA 190
         ::::::::::::   ::::::   |||:::|||         :::
     320 leGluSerThrSerThrThrPheThrSerThrAlaSerThrSerThrSer 336

     191 TCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAG 240
         |||          |||   ::::::   |||      |||   |||   ||
     337 SerThrSerThrThrGlnGlnSerSerSerThrIleThrSerSerProSe 353

     241 CAGGCGGCCATTAAAG.........................GAAATAT 263
         |         |||:::                          ||||||:
```

```
 353 rSerThrThrLeuSerThrSerIleProThrThrThrThrProGluIleT 370

 264 CGGCTACATTGTCCGCTTTTCCGATCACGGGC..................... 295
     :::::|||||||||:::::::|||   ::::::
 370 hrSerThrLeuSerSerLeuProAspAsnAlaIleCysSerTyrLeuAsp 386

 296 ......ACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGAT 339
           |||   |||          :::|||||||||
 387 GluThrThrThrSerThrThrPheThrThrThrMetLeuThrSerThrTh 403

 340 TCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCA 389
           :::|||   :::      |||   :::      |||::::::
 403 rThrGluGluProSerThrSerThrThrThrThrGluValThrSerThrS 420

 390 TTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGG 439
       :::|||   |||:::|||   |||   |||:::::::|||
 420 erSerThrValThrThrThrGluProThrThrThrLeuThrThrSerThr 436

 440 GCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGAC 489
     |||:::                                  |||:::|||||
 437 AlaSer................................ThrSerThrTh 442

 490 ATAAAAGGCGTTGCCCAAAATA......TCCGCCTCAACCTGACCGACAA 533
     |      :::      :::      |||   ||||||   |||:::|
 442 rGluProSerThrSerThrValThrThrSerProSerThrSerProValT 459

 534 CCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTA 583
     ||::::::      ::::::   |||:::   :::|||:::   :::
 459 hrSerThrValThrSerSerSerSerSerSerThrThrValThrThrPro 475

 584 TGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGC 633
                       :::|||:::|||::::::         :::::
 476 Thr..............SerThrGluSerThrSerThrSerProSerSe 487

 634 CCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGC 683
     :         |||:::   :::   |||         :::|||:::   :
 487 rThrValThrThrSerThrThrAlaProSerThrSerThrThrGlyProS 504

 684 AGATATCGTCAAAAACA...TCATCGGCGCGGCAGGAGAAATTGTCGGCG 730
     ::   ||||||:::|||   ||||||:::   :::::::::::::|||:::
 504 erSerSerSerSerThrProSerSerThrAlaSerSerSerValSerSer 520

 731 CAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATG 780
        |||
 521 ThrAla..................................... 522

 781 CACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGA 830
                                             :::|||||||
 523 .....................................SerSerThr. 525

 831 TTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCG 880
                       :::|||:::|||      ||||||   |||:::
 526 ................GlnSerSerThrSerThrGlnGlnSerSerThr 536

 881 ATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGC 930
           :::|||:::|||                  :::   :::|||:::
 537 ThrThrLysSerGluThrThrThrSerSerAspGlyThrAsnProAspPh 553

 931 AATATCTTTACGGCAGTCATCCCCG....................... 955
           :::::::::::::::
 553 eTyrPheValGluLysAlaThrThrThrPheTyrAspSerThrSerValA 570

 956 ..........TCAAAGGGATTGGAGCTGTTCGGG.............. 979
                 ::::::||||||   ::::::|||
 570 snLeuThrLeuAsnSerGlyLeuGlyIleIleGlyTyrGlnThrSerIle 586

 980 GAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGC......GG 1023
     |||   |||:::   ::::::|||:::::::::   ::::::      ||
 587 GluCysThrSerProThrSerSerAsnTyrValSerThrThrLysAspGl 603

1024 TCGCAGATGGGCGAGATCGCATTGCCGAAAG.........GGAAATCCGC 1064
     |      ::::::::::|||                     |||:::   |
 603 yAlaCysPheThrLysSerValSerMetProArgLeuGlyGlyThrTyrP 620

1065 CGTCAGCGACAATTTTG..............CCGATGCGGCATACGCCA 1099
     ||:::::::|||::::::                   :::|||   |||
 620 roAlaSerThrPheValGlyProGlyAsnTyrThrPheArgAlaThrMet 636

1100 AATACCCGTCCC............CTTACCATTCCCGAAATATCCGTTCA 1137
     :::|||                     |||      ::::::   :::||
 637 ThrThrAspAspLysLysValTyrTyrThrTyrAlaAsnValTyrIleGl 653

1138 AACTTGGAGCAGCGTTACGGCAAAGAAAACATCA 1171
     |   :::|||||||:::|||   ::::::   |||
 653 nGluTyrSerSerThrThrIleGluSerGluSer 664
```

```
seq_name: SwissProt_40:A180_RAT

seq_documentation_block:
ID   A180_RAT       STANDARD;      PRT;   915 AA.
AC   Q05140;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Clathrin coat assembly protein AP180 (Clathrin coat associated protein
DE   AP180).
GN   SNAP91.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=93178442; PubMed=8440257;
RA   Morris S.A., Schroeder S., Plessmann U., Weber K., Ungewickell E.;
RT   "Clathrin assembly protein AP180: primary structure, domain
RT   organization and identification of a clathrin binding site.";
RL   EMBO J. 12:667-675(1993).
CC   -!- FUNCTION: ADAPTINS ARE COMPONENTS OF THE ADAPTOR COMPLEXES WHICH
CC       LINK CLATHRIN TO RECEPTORS IN COATED VESICLES. CLATHRIN-
CC       ASSOCIATED PROTEIN COMPLEXES ARE BELIEVED TO INTERACT WITH THE
CC       CYTOPLASMIC TAILS OF MEMBRANE PROTEINS, LEADING TO THEIR SELECTION
CC       AND CONCENTRATION. BINDING OF AP180 TO CLATHRIN TRISKELIA INDUCES
CC       THEIR ASSEMBLY INTO 60-70 NM COATS.
CC   -!- SUBCELLULAR LOCATION: COMPONENT OF THE COAT SURROUNDING THE
CC       CYTOPLASMIC FACE OF COATED VESICLES IN THE PLASMA MEMBRANE.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; A LONG FORM (SHOWN HERE) AND A
CC       SHORT FORM; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- DOMAIN: POSSESSES A THREE DOMAIN STRUCTURE: THE N-TERMINAL 300
CC       RESIDUES HARBOR A CLATHRIN BINDING SITE, AN ACIDIC MIDDLE DOMAIN
CC       450 RESIDUES, INTERRUPTED BY AN ALA-RICH SEGMENT, AND THE C-
CC       TERMINAL DOMAIN (166 RESIDUES).
CC   -!- PTM: PHOSPHORYLATED (BY SIMILARITY).
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; X68877; CAA48748.1; -.
DR   EMBL; X68878; CAA48749.1; -.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR001026; ENTH.
DR   Pfam; PF01417; ENTH; 1.
DR   SMART; SM00273; ENTH; 1.
KW   Coated pits; Alternative splicing; Phosphorylation.
FT   DOMAIN      410    413       POLY-THR.
FT   DOMAIN      535    539       POLY-ALA.
FT   DOMAIN      547    550       POLY-ALA.
FT   DOMAIN      678    683       POLY-SER.
FT   DOMAIN      723    729       POLY-SER.
FT   VARSPLIC    614    632       MISSING (IN SHORT ISOFORM).
```

SQ SEQUENCE 915 AA; 93518 MW; 32EC1B38EC5DF8C0 CRC64;

alignment_scores:
Quality: 119.00 Length: 553
Ratio: 0.500 Gaps: 27
Percent Similarity: 43.038 Percent Identity: 22.423

alignment_block:
US-09-303-518D-465 x A180_RAT ..

Align seg 1/1 to: A180_RAT from: 1 to: 915

```
 107 TTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGC 156
     |||:::|||:::|||   :::   |||:::::::::::::      |||::
 331 PheAlaThrAlaSerAlaAlaAlaProValSerSerAlaLysProSerSe 347

 157 A.........................GCAGGGGGGAACTTGCCGAGCG 179
     :                          :::|||::::::     :::|
 347 rAspLeuLeuAspLeuGlnProAspPheSerGlyAlaArgAlaGlyAlaA 364

 180 CAGCGGTCATATCGGATTGGGAAACATACAAAGCCATCAGTTGG..... 223
     |||||   :::          |||      |||   :::|||
 364 laAlaProValProPro...ProThrGlyGlyAlaThrAlaTrpGlyAsp 379

 224 .....................GCAACCTGTTCATCC......AGCAGGCG 246
                          |||:::   ||||||      |||:::
 380 LeuLeuGlyGluAspSerLeuAlaAlaLeuSerSerValProSerGluAl 396

 247 GCCATTAAAGGAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCA 296
      |||:::::::::      |||          :::   |||   |||   |
 396 aProIleSerAspProPheAlaProGluProSerProProThrThrThrT 413

 297 CGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATG 346
     ||:::            |||::::::
 413 hrGluProAlaSerAlaSerAlaSerAlaThrThrAlaValThrAlaAla 429

 347 AAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGAC 396
               |||   |||      :::|||||||:::||||||   |||
 430 ThrThrGluValAspLeuPheGlyAspAlaPheAlaAlaSerProGlyGl 446

 397 GGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGG 446
                        |||:::|||      |||      |||   |
 446 u..............AlaProAlaAlaSerGluGlyAlaThrAlaProA 458

 447 CTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAG 496
     ||   |||   |||   |||   ::::::   :::::::::      .
 458 laThrProAlaProValAlaAlaAlaLeuAspAlaCysSer......... 471

 497 GCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGA 546
     :::          ::::::   |||        :::   |||||||||::
 472 GlyAsnAspProPheAlaProSerGluGlySerAlaGluAlaAlaProGl 488

 547 CAACGGCTTGTCGACCGTT............TCCACAATACCGGTAGTAT 584
     :      |||   ::::::            ::::::   |||||||||
 488 uLeuAspLeuPheAlaMetLysProProGluThrSerAlaProValValT 505

 585 GCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCA......CCCGAT 628
                         |||   |||:::||||||      |||
 505 hr..............ProThrAlaSerThrAlaProProValProAla 516

 629 ACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACGGC 678
     ||||||||||||   :::      |||:::             :::|||::
 517 ThrAlaProSerProAlaProThrAlaValAlaAlaAlaThrAlaAlaThrTh 533

 679 ACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGG 728
     :      ::::::   |||:::::::|||      :::      :::|
 533 rThrAlaAlaAlaAlaAlaThrThrThrAlaThrThrSerAlaAlaAlaA 550

 729 CGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTA 778
     ||                         :::   |||          |||
 550 laThr...................ThrAlaAlaAlaProProAlaLeu 559

 779 TGCACGGCTTGGGTCTGCTTTCCAC.....................C 804
           :::|||:::   |||
 560 AspIle.PheGlyAspLeuPheAspSerAlaProGluValAlaAlaAlaS 576

 805 GAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTCAA 854
       |||   |||   |||         :::      :::|||:::
 576 erLysProAspValAla...ProSerIleAspLeuPheGlyThrAsp... 590

 855 AGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCCAAAACCCCAA.. 902
         :::   ||||||   ||||||   :::||||||   |||:::
 591 ......AlaPheSerSerProProArgGlyAlaSerProValProGluSe 605

 903 .....................................TGCCGCACAA 912
                                          ||| |||:::
 605 rSerLeuThrAlaAspLeuLeuSerGlySerGlyPheHisCys.AlaGlu 621

 913 GGCATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGG 962
           :::   |||      :::||||||||||||           ::
 622 AspAspArgHisValProLeuPhePheThrAlaValAspAlaPheAlaAl 638

 963 GATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCACGGCACATC 1012
     :   :::   :::                                |||||   |
 638 aProSerProAlaSer.....................ThrAlaSerP 647

1013 CTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCC 1062
     ||:::|||   ::::::   |||   |||   |||
 647 roAlaLysAlaGluSerSerGlyValIleAspLeu.............. 658

1063 GCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCC 1112
           :::|||   |||:::::::::|||           |||:::||
 659 ...PheGlyAspAlaPheGlySerSerAlaSerGluThrGlnProAlaPr 674

1113 T......TACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACG 1156
     |            |||   :::      :::::::|||         :::|
 674 oGlnAlaValSerSerSerSerAlaSerAlaAspLeuLeuAlaGlyPheG 691

1157 GCAAAGAAAACATCACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAAT 1206
     ||   :::   ::::::   |||||||:::|||   :::   ::::::|||
 691 lyGlySerPheMetAlaProSerThrThrProValThrProAlaGlnAsn 707

1207 GTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGG 1256
                |||      :::|||:::   ::::::   |||
 708 ............AsnLeuLeuGlnProAsnPheGluAlaAlaPheGlyTh 720

1257 TAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTA 1306
               |||:::   :::::::::   :::::::|||        :
 720 rThr......ProSerThrSerSerSerSerSerPheAspProSerGlyA 735

1307 ATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGTCTTTAATCCT 1356
     ::      :::|||   :::   |||      :::|||
 735 spLeuLeuMetProThrMetAlaProSerGlyGlnPro............ 747

1357 AAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATTCA 1406

 747 ................................................. 747

1407 ATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCTAAAA 1456
                          :::   :::      .:::||||||
 748 ....................AlaProValSerMetValProPro.... 755

1457 ATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAATGGATATTTG 1506
           |||||||:::           ||||||   ::::::      |||
 756 .....SerProAlaMetSerAlaSerLysGlyLeuGlySerAsp...Leu 769

1507 GATAAA..............TTTGGTAATGAATGGACTAAAGGTCCATC 1541
     |||:::                |||||          :::|||   ::
 770 AspSerSerLeuAlaSerLeuValGlyAsnLeuGlyIleSerGlyThrTh 786

1542 AAGAACTAAAGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAA...A 1588
     :      ||||||   ::::::::::|||::::::      :::|||   |
 786 rSerLysLysGly...AspLeuGlnTrpAsnAlaGlyGluLysLysLeuT 802

1589 CAGGA 1593
```

```
        |||||
    802 hrGly 803

seq_name: SwissProt_40:NRG3_MOUSE

seq_documentation_block:
ID   NRG3_MOUSE     STANDARD;      PRT;   713 AA.
AC   O35181;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Pro-neuregulin-3 precursor (Pro-NRG3) [Contains: Neuregulin-3 (NRG-
DE   3)].
GN   NRG3.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=97420720; PubMed=9275162;
RA   Zhang D., Sliwkowski M.X., Mark M., Frantz G., Akita R., Sun Y.,
RA   Hillan K., Crowley C., Brush J., Godowski P.J.;
RT   "Neuregulin-3 (NRG3): a novel neural tissue-enriched protein that
RT   binds and activates ErbB4.";
RL   Proc. Natl. Acad. Sci. U.S.A. 94:9562-9567(1997).
CC   -!- FUNCTION: DIRECT LIGAND FOR THE ERBB4 TYROSINE KINASE RECEPTOR.
CC       BINDING RESULTS IN LIGAND-STIMULATED TYROSINE PHOSPHORYLATION AND
CC       ACTIVATION OF THE RECEPTOR. DOES NOT BIND TO THE EGF RECEPTOR,
CC       ERBB2 OR ERBB3 RECEPTORS.
CC   -!- SUBCELLULAR LOCATION: EXISTS AS AN TYPE I MEMBRANE PROTEIN AND AS
CC       A PROTEOLYTICALLY RELEASED SOLUBLE GROWTH FACTOR FORM. THE
CC       MEMBRANE-BOUND FORM DOES NOT SEEM TO BE ACTIVE (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: EXPRESSED IN SYMPATHETIC, MOTOR, AND SENSORY
CC       NEURONS.
CC   -!- DEVELOPMENTAL STAGE: DETECTED AS EARLY AS E11. IN E13 EMBRYOS,
CC       DETECTED MAINLY IN THE NERVOUS SYSTEM. IN E16 EMBRYOS, DETECTED IN
CC       THE BRAIN, SPINAL CORD, TRIGEMINAL, VESTIBULAR-COCHLEAR, AND
CC       SPINAL GANGLIA. IN ADULTS, EXPRESSED IN SPINAL CORD, AND NUMEROUS
CC       BRAIN REGIONS.
CC   -!- DOMAIN: THE CYTOPLASMIC DOMAIN MAY BE INVOLVED IN THE REGULATION
CC       OF TRAFFICKING AND PROTEOLYTIC PROCESSING. REGULATION OF THE
CC       PROTEOLYTIC PROCESSING INVOLVES INITIAL INTRACELLULAR DOMAIN
CC       DIMERIZATION (BY SIMILARITY).
CC   -!- DOMAIN: ERBB RECEPTOR BINDING IS ELICITED ENTIRELY BY THE EGF-LIKE
CC       DOMAIN (BY SIMILARITY).
CC   -!- PTM: PROTEOLYTIC CLEAVAGE CLOSE TO THE PLASMA MEMBRANE ON THE
CC       EXTERNAL FACE LEADS TO THE RELEASE OF THE SOLUBLE GROWTH FACTOR
CC       FORM (BY SIMILARITY).
CC   -!- PTM: EXTENSIVE GLYCOSYLATION PRECEDES THE PROTEOLYTIC CLEAVAGE (BY
CC       SIMILARITY).
CC   -!- SIMILARITY: CONTAINS 1 EGF-LIKE DOMAIN.
CC   -!- SIMILARITY: BELONGS TO THE NEUREGULIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF010130; AAB70914.1; -.
DR   MGD; MGI:1097165; Nrg3.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002154; Neuregulin.
DR   Pfam; PF00008; EGF; 1.
DR   Pfam; PF02158; Neuregulin; 1.
DR   SMART; SM00181; EGF; 1.
DR   PROSITE; PS00022; EGF_1; 1.
DR   PROSITE; PS01186; EGF_2; 1.
KW   Growth factor; EGF-like domain; Transmembrane; Multigene family.
FT   CHAIN         1    713       PRO-NEUREGULIN-3, MEMBRANE-BOUND FORM.
FT   CHAIN         1    361       NEUREGULIN-3.
FT   DOMAIN        1    362       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    363    383       INTERNAL SIGNAL SEQUENCE (POTENTIAL).
FT   DOMAIN      384    713       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN      105    287       SER/THR-RICH.
FT   DOMAIN      288    331       EGF-LIKE.
FT   DOMAIN       13     21       POLY-ALA.
FT   DOMAIN       26     34       POLY-ALA.
FT   DOMAIN      127    135       POLY-THR.
FT   DOMAIN      250    253       POLY-ALA.
FT   DOMAIN      254    263       POLY-SER.
FT   DOMAIN      264    267       POLY-THR.
FT   DISULFID    292    306       BY SIMILARITY.
FT   DISULFID    300    319       BY SIMILARITY.
FT   DISULFID    321    330       BY SIMILARITY.
SQ   SEQUENCE   713 AA;  77369 MW;  9F7D1D5E7FC8DCF0 CRC64;


alignment_scores:
           Quality:  118.50            Length:    554
             Ratio:  0.517               Gaps:     30
 Percent Similarity:  41.336   Percent Identity: 23.105

alignment_block:
 US-09-303-518D-465 x NRG3_MOUSE   ..

 Align seg 1/1  to: NRG3_MOUSE  from: 1  to: 713

     275 TCCGCTTTTCCG...ATCACGGGCACGAAGTCCATTCCCCCTTCGACAAC 321
         |||:::||||||            |||   :::         ||||||||
     119 SerSerPheProLysAlaMetGluThrThrThrThrThrSerThrTh 135

     322 CATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATT 371
         |   |||       |||                :::         |||   :
     135 rSerProAlaThrProSerAlaGlyGlyAlaAlaSerSerArgThrProA 152

     372 CAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACG 421
         ::   :::::::::   :::   |||   |||:::::::      ||||||
     152 snArgIleSerThrArgLeuThrThrIleThrArgAla......ProThr 166

     422 GCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGG 471
            :::   |||||||||         :::|||:::
     167 ArgPheProGlyHisArg.........ValProIle.............. 175

     472 GATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAA 521
                               :::|||   |||:::      ::::::|
     176 .....................ArgAlaSerProArg......SerThrT 183

     522 CCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACA 571
         ||      :::||||||||||||   :::   |||||||||       |||
     183 hrAlaArgAsnThrAlaAlaProProThrValLeuSerThr......Thr 197

     572 ATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCC 621
         |||                     :::   :::|||   ::::::   ||
     198 AlaProPhePhe............SerSerSerThrProGlySerArgPr 210

     622 ACCC..................................................... 625
         ||||
     210 oProMetProGlyAlaProSerThrGlnAlaMetProSerTrpProThrA 227

     626 ............................GATACAGCCCCGAGCTGGACA 646
                                    |||:::::|||||||||||||
     227 laAlaTyrAlaThrSerSerTyrLeuHisAspSerThrProSerTrpThr 243

     646 ..................................................... 646

     244 LeuSerProPheGlnAspAlaAlaAlaAlaSerSerSerSerProSerSe 260

     647 .......GATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATAT 689
                 :::   :::      |||:::   |||||||:::   ::::::
     260 rThrSerSerThrThrThrThrProGluThrSerThrSerProLysPheH 277
```

```
 690 CGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATG 739
     :::   |||   |||::::::::::         |||   :::
 277 isThrThrThrTyrSerThrGluArg.........SerGluHisPheLys 290

 740 CCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGC...ACGGC 786
     |||||||||              :::   |||      |||   :::
 291 ProCysArg............AspLysAspLeuAlaTyrCysLeuAsnAs 303

 787 TTGGGTCTGCTTTCCACCGAAAACA..................AGATGGC 818
         ||||||      :::|||                  :::   |
 303 pGlyGluCysPheValIleGluThrLeuThrGlySerHisLysHisCysA 320

 819 GCGCATCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAG 868
     ||::::::      :::|||      |||   ::::::   :::|||:::
 320 rgCysLysGluGlyTyrGlnGlyValArgCysAspGlnPheLeuProLys 336

 869 CAGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAA.GGCAT 917
         |||   :::::::::         |||      |||   |||||
 337 ThrAspSerIleLeuSerAsp.........ProThrAspHisLeuGlyIl 350

 918 AGAA.....................................GCCGTCA 928
     ||||                                    ::::::|
 350 eGluPheMetGluSerGluAspValTyrGlnArgGlnValLeuSerIleS 367

 929 GCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTT... 975
     ||   |||:::   :::::::::  |||      :::   |||:::
 367 erCysIleIlePheGlyIleValIleValGlyMetPheCysAlaAlaPhe 383

 976 ......CGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAA 1019
           :::::::|||              |||      |||   :::||
 384 TyrPheLysSerLysLysGlnAlaLysGlnIleGlnGluHis...LeuLy 399

1020 GCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCA 1069
     |:::|||||   |||:::           |||:::::|||:::   |
 399 sGluSerGlnAsnGlyLysAsnTyrSerLeuLysAlaSerSerThrLysS 416

1070 GCGACAATTTT.................................GCC 1083
     ||:::::::::                                  |||
 416 erGluSerLeuMetLysSerHisValHisLeuGlnAsnTyrSerLysAla 432

1084 GATGCGGCAT.........................ACGCCAAATA 1103
     ||| ||||||                        :::      :
 433 Asp.ArgHisProValThrAlaLeuGluLysIleMetGluSerSerPheS.449

1104 CCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTT 1153
     ::   |||   :::::|||||:::::          :::|||
 449 erAlaProGlnSerPheProGluValThrSerProAspArgGlySerGln 465

1154 ACGGCAAAGAAAACATCACCTCCTCAA.........CCGTGCCGCCGTCA 1194
           |||      ||||||   |||                ::::|::
 466 ProIleLysHisHisSerProGlyGlnArgSerGlyMetLeuHisArgAs 482

1195 AACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGT 1244
     :|||   |||                ::::|||         |||:::
 482 nThrPheArg.................ArgAlaProProSerProArgS 493

1245 GCCGTTTG...ACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAAT 1291
       ||||||      :::   |||
 493 erArgLeuGlyGlyIleValGly.......................... 500

1292 ACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAA 1341
                     |||   |||::::::
 501 .................ProAlaTyrGlnGln................. 505

1342 CCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTAT 1391
                        |||::::::      |||
 506 ....................LeuGluGluSerArgIle.......... 511

1392 AACTGCCAGAATTCAATACGCAAAATTACCAAGGC...AAGGTAGAATCA 1438
         |||:::   :::|||          ||||||   :::|||::::::
 512 .....ProAspGlnAspThrIleProCysGlnGlyIleGluValArgLys 526

1439 GATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGA 1488
```

```
          ||||||||||||   |||   |||         :::         ||
 527 ThrIleSerHisLeuProIleGlnLeuTrpCysValGluArgProLeuAs 543

1489 CCTAATAATGGATATTTGGATAAATTTGGTAATGAATGGACTAAAGGTCC 1538
          ||||                  :::   |||   |||||||||:::::::
 543 pLeu..................LysTyrValSerAsnGlyLeuArgThrG 554

1539 ATCAA..................GAACTAAAGGTCAAGAATTTGAATGGG 1570
          ::|||                :::|||      ::::::   |||
 554 lnGlnAsnAlaSerIleAsnMetGlnLeuProSerArgGluThrAsnPro 570

1571 ATGTTCAATTGTCTAAAACAGGAAGAGAGCAACTTGGATGGGCTAGTAGG 1620
        ||||||   |||   |||::::::                ||||||||
 571 TyrPheAsnSerLeuAspGlnLysAsp..............LeuValGl 582

1621 GATGGTAAGC 1630
     |   :::|||
 582 yTyrLeuSer 585
```

```
seq_name: SwissProt_40:FYB_MOUSE

seq_documentation_block:
ID   FYB_MOUSE      STANDARD;      PRT;   819 AA.
AC   O35601; Q9Z2H3;
DT   15-JUL-1999 (Rel. 38, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   FYN-binding protein (FYN-T-binding protein) (FYB-120/130) (p120/p130)
DE   (SLP-76 associated phosphoprotein) (SLAP-130).
GN   FYB.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A. (ISOFORM FYB-120).
RC   TISSUE=T-cell lymphoma;
RX   MEDLINE=97352826; PubMed=9207119;
RA   da Silva A.J., Li Z., de Vera C., Canto E., Findell P., Rudd C.E.;
RT   "Cloning of a novel T-cell protein FYB that binds FYN and SH2-domain-
RT   containing leukocyte protein 76 and modulates interleukin 2
RT   production.";
RL   Proc. Natl. Acad. Sci. U.S.A. 94:7493-7498(1997).
RN   [2]
RP   SEQUENCE FROM N.A. (ISOFORM FYB-130).
RC   TISSUE=Hybridoma;
RX   MEDLINE=99428514; PubMed=10497204;
RA   Veale M., Raab M., Li Z., da Silva A.J., Kraeft S.-K., Weremowicz S.,
RA   Morton C.C., Rudd C.E.;
RT   "Novel isoform of lymphoid adaptor FYN-T-binding protein (FYB-130)
RT   interacts with SLP-76 and up-regulates interleukin 2 production.";
RL   J. Biol. Chem. 274:28427-28435(1999).
CC   -!- FUNCTION: ACTS AS A ADAPTER PROTEIN OF THE FYN AND SH2-DOMAIN-
CC       CONTAINING LEUCOCYTE PROTEIN-76 (SLP76) SIGNALING CASCADES IN T
CC       CELLS. MODULATES THE EXPRESSION OF INTERLEUKIN-2 (IL-2).
CC   -!- SUBUNIT: INTERACTS WITH FYN AND SLP76.
CC   -!- SUBCELLULAR LOCATION: Nuclear and cytoplasmic.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; FYB-130 (SHOWN HERE) AND FYB-
CC       120; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: EXPRESSED IN HEMATOPOIETIC TISSUES SUCH AS
CC       MYELOID AND T CELLS, SPLEEN AND THYMUS. NOT EXPRESSED IN B CELLS,
CC       NOR IN NON-LYMPHOID TISSUES. FYB-130 IS PREFERENTIALLY EXPRESSED
CC       IN MATURE T-CELLS COMPARED TO FYB-120, WHEREAS THYMOCYTES SHOWED A
CC       GREATER RELATIVE AMOUNT OF FYB-120.
CC   -!- PTM: T-CELL RECEPTOR LIGATION LEADS TO INCREASED TYROSINE
CC       PHOSPHORYLATION.
CC   -!- SIMILARITY: CONTAINS 1 SH3 DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
```

```
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ------------------------------------------------------------------------
DR   EMBL; AF001863; AAB62227.1; -.
DR   EMBL; AF061744; AAD03267.1; -.
DR   MGD; MGI:1346327; Fyb.
DR   InterPro; IPR001452; SH3.
DR   Pfam; PF00018; SH3; 1.
DR   SMART; SM00326; SH3; 1.
DR   PROSITE; PS50002; SH3; 1.
KW   SH3 domain; Phosphorylation; Nuclear protein; Coiled coil;
KW   Alternative splicing.
FT   DOMAIN      343    348       POLY-PRO.
FT   DOMAIN      381    385       POLY-PRO.
FT   DOMAIN      448    495       COILED COIL (POTENTIAL).
FT   DOMAIN      479    493       NUCLEAR LOCALIZATION SIGNAL (POTENTIAL).
FT   DOMAIN      584    587       SH2-BINDING (TO SLP-76).
FT   DOMAIN      606    609       POLY-PRO.
FT   DOMAIN      615    618       SH2-BINDING (TO FYN).
FT   DOMAIN      622    625       POLY-ASP.
FT   DOMAIN      696    699       POLY-PRO.
FT   DOMAIN      699    756       SH3.
FT   DOMAIN      710    736       NUCLEAR LOCALIZATION SIGNAL (POTENTIAL).
FT   DOMAIN      739    805       SH3-LIKE.
FT   VARSPLIC    627    672       MISSING (IN ISOFORM FYB-120).
SQ   SEQUENCE   819 AA;  90055 MW;  6222CC7EF1CA2BBD CRC64;


alignment_scores:
            Quality:  118.50              Length:    568
              Ratio:  0.502                Gaps:     28
 Percent Similarity:  41.549   Percent Identity: 19.542

alignment_block:
 US-09-303-518D-465 x FYB_MOUSE   ..

 Align seg 1/1  to: FYB_MOUSE  from: 1  to: 819

     125 TCGAACCCG..............ACGGGAAATACCACCTATTCGGCAGC 159
         |||||||||              |||::::::      ::::::::::||
       8 SerAsnProThrGluGluAlaAlaThrSerSerArgProPheLysValAl 24

     160 AGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACA 209
         ||||   :::   ||||||:::
      24 aGlyGlnSerSerProSerGlyIle........................ 32

     210 AAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAA 259
                                 :::||||||:::   |||   :::
      33 .........................GlnSerArgLysAsnLeuPheAsp 40

     260 ATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCC 309
           ::::::      :::|||   |||   :::|||   :::|||:::||
       41 AsnGlnGly......AsnAlaSerProProAlaGlyProSerSerMetPr 55

     310 CCCT...TCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAG 356
         |      :::||||||   |||      |||:::      ||||||:::
      55 oLysPheGlyThrThrLysPro......ProLeuAlaAlaLysProThrT 70

     357 TCCCG.................................TTGACGGATTCA 373
                                              |||      :::
      70 yrGluGluLysProGluLysGluProLysProProPheLeuLysProThr 86

     374 GCCTTTACCGCATCCATTGGGACGGATACGAACACCATC.......... 412
         :::   :::      :::||||||:::   |||:::::
      87 GlyGlySerProArgPheGlyThrGlnProAsnSerValSerArgAspPr 103

     413 ....................................CCGCCGACGGCTA 425
                                              ||||||::::
     103 oGluValLysValGlyPheLeuLysProValSerProLysProThrSerL 120

     426 TGACGGGCCACAGGGCGGCGGCTATCCCGCTC...CCAAAGGCGCGAGGG 472
         ::|||   :::   :::      :::   |||   |||   ::::::
     120 euThrLysGluAspSerLysProValValLeuArgProProGlyAsnLys 136

     473 ATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAAC 522
         :::::::::   :::      :::   |||         :::   :::
     137 LeuHisAsnLeuAsnGlnGluSerAspLeuLysThrProGlyProLysPr 153

     523 CTGACCGACAACCGCAGCACCGGACAAC............GGCTTGTCGA 560
           |||:::      :::|||:::|||            |||:::|||
     153 oGlyProAlaProProValProGluAsnGluLeuLysProGlyPheSerL 170

     561 CCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGAT 610
           |||
     170 ysVal.......................................... 171

     611 TCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCC 660
         :::::::|||                |||:::      |||
     172 AlaGlyAlaLysSerLysPheMetProAlaAlaGlnAspThrAspSerLy 188

     661 GCCGAAGCTT......TCAACGGCACTGCAGATATCGTCAAAAACATCAT 704
          |||::::::         |||      |||      |||   :::::::
     188 sProArgPheProArgHisThrPheGlyGlnLysProSerLeuSerThrG 205

     705 CGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAA 754
         ::      |||||||:::   :::::::::   :::|||   :::
     205 luAspSerGlnGluGluAsnThrSerLysAsnValProValGln...... 219

     755 GCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACC 804

     219 ..................................................... 219

     805 GAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAACTCAA 854
                       :::::|||   :::   |||:::      :::||||
     220 .............LysGlySerProVal...GlnLeuGlyAlaLysSerL 231

     855 AGACTATGCCGCAGCAGCCA.....................TCCGCGATT 883
         ||      |||   :::|||                        :::
     231 ysGlyAlaProPheLysProProLysGluAspProGluAspLysAspHis 247

     884 GGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAAT 933
         |||      ::::::|||:::|||      :::   |||||||:::|||
     248 GlyAlaProSerSerProProProGlyValValLeuLysProAlaAla.. 263

     934 ATCTTTACGGCAGTCATCCCCG......TCAAAGGGATTGGAGCTGTTCG 977
                 ::::::||||||      ||||||::::::|||
     264 .......SerArgGlySerProGlyLeuSerLysAsnPheGluGluLysL 278

     978 GGGAAAATACGGCTTGGGCGGCATCACGGCACATC.............. 1012
           |||:::                :::::::::|||
     278 ysGluAspArgLysThrAspLeuAlaLysAsnIlePheLeuAsnLysLeu 294

    1013 .....................................CTGTCAAG 1020
                                              |||:::::
     295 AsnGlnGluGluProAlaArgPheProLysAlaProSerLysLeuThrAl 311

    1021 CGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCAG 1070
         :|||      |||::::::|||:::   ::::::|||:::      ||||
     311 aGlyThrProTrpGlyGlnSerGlnGluLysGluGlyAspLysAsnSerA 328

    1071 CGACAATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCATT 1120
         ||                        ||||||:::   :::   |||
     328 la....................ThrProLysGlnLysAlaLeu...... 335

    1121 CCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATC 1170
         |||   :::||||||                                  ::
     336 ProProLeuSerVal.............................LeuGl 342

    1171 ACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAA 1220
         :||||||   |||      |||::::::      ||||||   |||:::::
     342 yProProProPro..LysProAsnArgProProAsnValAspLeuThrAr 358

    1221 CAAACGCCAC.....................CCGAAGACCAAAGTGCCGT 1249
         :   |||                          |||:::::::::|||:
     358 gPheArgLysAlaAspSerAlaAsnSerAlaThrLysSerGlnThrProT 375
```

```
1250 TTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACG 1299
      :::::      ::::::|||
 375 yrSerThrThrSerLeuPro..............................  381

1300 AGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGTCTT 1349
                     |||   ::::::|||   ::::::|||
 382 ............ProProProProThrHisProAlaSerGlnPro.....  392

1350 TAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCA 1399
                   :::   ::::::|||
 393 ............ProLeuProAlaSerHis.....................  398

1400 GAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCA 1449
                        |||   :::      :::      :::|||
 399 ...................ProAlaHisProProValProSerLeuPro  408

1450 CCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAAT.. 1497
     |||:::|||   :::|||         |||   |||   |||   |||
 409 ProArgAsnIleLysProProLeuAspLeu...LysHisProIleAsnAs  424

1498 ............GGATATTTGGATAAATTTGGTAATGAATGGACTAAAG 1534
                 |||   :::   :::   |||:::       :::
 424 pGluAsnGlnAspGlyValMetHisSerAspGlyThrGlyAsnLeuGluG  441

1535 GTCCATCAAGAACTAAAGGTCAAGAATTTGAATGGGATGTTCAATTGTCT 1584
           :::::::::::|||:::   :::|||   |||:::::   |||
 441 luGluGlnGluSerGluGlyGluThrTyrGlu...AspIleAspSerSer  456

1585 AAA 1587
     |||
 457 Lys 457
```

seq_name: SwissProt_40:CAPP_MYCLE

seq_documentation_block:

```
ID   CAPP_MYCLE     STANDARD;      PRT;   934 AA.
AC   P46710; Q9CCN5;
DT   01-NOV-1995 (Rel. 32, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Phosphoenolpyruvate carboxylase (EC 4.1.1.31) (PEPCASE) (PEPC).
GN   PPC OR ML0578 OR B1496_C3_207.
OS   Mycobacterium leprae.
OC   Bacteria; Firmicutes; Actinobacteria; Actinobacteridae;
OC   Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium.
OX   NCBI_TaxID=1769;
RN   [1]
RP   SEQUENCE FROM N.A.
RA   Smith D.R., Robison K.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=TN;
RX   MEDLINE=21128732; PubMed=11234002;
RA   Cole S.T., Eiglmeier K., Parkhill J., James K.D., Thomson N.R.,
RA   Wheeler P.R., Honore N., Garnier T., Churcher C., Harris D.,
RA   Mungall K., Basham D., Brown D., Chillingworth T., Connor R.,
RA   Davies R.M., Devlin K., Duthoy S., Feltwell T., Fraser A., Hamlin N.,
RA   Holroyd S., Hornsby T., Jagels K., Lacroix C., Maclean J., Moule S.,
RA   Murphy L., Oliver K., Quail M.A., Rajandream M.A., Rutherford K.M.,
RA   Rutter S., Seeger K., Simon S., Simmonds M., Skelton J., Squares R.,
RA   Squares S., Stevens K., Taylor K., Whitehead S., Woodward J.R.,
RA   Barrell B.G.;
RT   "Massive gene decay in the leprosy bacillus.";
RL   Nature 409:1007-1011(2001).
CC   -!- FUNCTION: TO FORM OXALOACETATE, A FOUR-CARBON DICARBOXYLIC ACID
CC       SOURCE FOR THE TRICARBOXYLIC ACID CYCLE (BY SIMILARITY).
CC   -!- CATALYTIC ACTIVITY: Phosphate + oxaloacetate = H(2)O +
CC       phosphoenolpyruvate + CO(2).
CC   -!- PATHWAY: TRICARBOXYLIC ACID CYCLE.
CC   -!- SUBUNIT: HOMOTETRAMER (BY SIMILARITY).
CC   -!- SIMILARITY: BELONGS TO THE PEPCASE FAMILY.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; U00013; AAA17132.1; ALT_INIT.
DR   EMBL; AL583919; CAC30086.1; -.
DR   Leproma; ML0578; -.
DR   HSSP; P00864; 1FIY.
DR   InterPro; IPR001449; PEPcase.
DR   Pfam; PF00311; PEPcase; 2.
DR   PRINTS; PR00150; PEPCARBXLASE.
DR   PROSITE; PS00393; PEPCASE_2; 1.
DR   PROSITE; PS00781; PEPCASE_1; 1.
KW   Lyase; Carbon dioxide fixation; Allosteric enzyme;
KW   Tricarboxylic acid cycle; Complete proteome.
FT   ACT_SITE    161    161       BY SIMILARITY.
FT   ACT_SITE    594    594       BY SIMILARITY.
SQ   SEQUENCE   934 AA;  102515 MW;  3EE8FD762EC45180 CRC64;
```

alignment_scores:

```
            Quality:  118.00              Length:    580
              Ratio:  0.474                 Gaps:     31
 Percent Similarity:  42.931   Percent Identity: 21.897
```

alignment_block:

US-09-303-518D-465 x CAPP_MYCLE ..

Align seg 1/1 to: CAPP_MYCLE from: 1 to: 934

```
  58 ATGCATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGT 107
     ||||||         |||:::|||  :::|||:::  ::::::  ::
 446 MetHisGlnAspTyrSerSerLeuProGluAspGlnArgValLysLeuLe 462

 108 TCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCA 157
     ::::              |||      ::::::      |||  |||:
 462 uVal............AlaGluLeuGlyAsnArgArgProLeuValGlyA 475

 158 GCAGGGGGGAACTTGCCGAGCGCAGC...GGTCATATCGGATTGGGAAAC 204
     ::|||:::::::|||::::::  :::  |||:::::::::::::
 475 spArgAlaGlnLeuSerAspLeuAlaArgGlyGluLeuAlaValLeuAla 491

 205 ATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAA 254
           :::|||  :::          :::|||:::      |||
 492 AlaAlaAlaHisAlaVal.GluLeuTyrGlySerAlaAlaValProAsnT 508

 255 AGGAAATATCGGCTACATTGTCC.....................GCTTTT 283
           ||||||        |||                    |||:::
 508 yrIleIleSerMetCysGlnSerValSerAspValLeuGluValAlaIle 524

 284 CCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACAT 333
        :::      |||      :::      |||   ::::::|||:::
 525 LeuLeuLysGluThrGlyLeuLeuAspAlaSerGlySerGlnProTyrCy 541

 334 TCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGC........ 375
      |||:::   ::::::|||:::            |||:::
 541 sProValGlyIleSerProLeuPheGluThrIleAspAspLeuHisAsnG 558

 376 ................................CTTTACCGCATC... 387
                                     |||||||||
 558 lyAlaAlaIleLeuHisAlaMetLeuGluLeuProLeuTyrArgThrLeu 574

 388 ..............CATTGGGAC............GGATACGAACACCA 410
                  :::|||:::             ||||||:::
 575 ValAlaAlaArgGlyAsnTrpGlnGluValMetLeuGlyTyrSer..... 589

 411 TCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCA 460
                   |||:::   :::   |||||||||   |||   :
 590 ...............AspSerAsnLysAspGlyGlyTyrLeuAlaAlaA 601
```

```
 461 AAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAAT 510
     ::   |||      :::|||                          .:::
 601 snTrpAla......ValTyr..........................Arg 606

 511 ATCCGCCTCAACCTGACCGAC...AACCGCAGCACCGGACAACGGCTTGT 557
       :::|||   |||:::|||      |||:::||||||   ||||||
 607 AlaGluLeuAlaLeuValAspValAlaArgLysThrGlyIleArgLeu.. 622

 558 CGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACG 607
           ||||||:::   |||:::   :::   :::|||   |||   :
 623 .ArgLeuPheHisGlyArgGlyGlyThrValGlyArgGlyGlyGlyProS 639

 608 GATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAAT 657
     ::::::::                |||
 639 erTyrGlnAlaIleLeuAlaGlnProPro..................... 648

 658 GCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGG 707
           |||   ||||||:::      ::::::::::::
 649 ......GlyAlaValAsnGlySerLeuArgLeuThrGluGln........ 660

 708 CGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCG 757
           ||||||::::::::::|||      |||         |||::::
 661 .......GlyGluValIleAlaAlaLysTyrAlaGluProGlnIleAlaA 675

 758 AAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAA 807
     ::   :::           :::::::::|||         :::   |||
 675 rgArgAsn............LeuGluSerLeuValAlaAlaThrLeuGlu 687

 808 AACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAA........ 849
     :::   :::      ::::::   |||:::|||   |||:::
 688 SerThrLeuLeuAspValGluGlyLeuGlyAspAlaAlaGluSerAlaTy 704

 850 .......CTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCA.... 888
           |||   :::   |||:::   |||   |||::::::|||
 704 rAlaIleLeuAspGluValAlaGlyLeuAlaArgArgSerTyrAlaGluL 721

 889 ..GTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATC 936
       |||::::::|||:::   :::                   :::
 721 euValAsnThrProGlyPheVal.....................AspTyr 730

 937 TTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGG........ 978
     |||   |||      ||||||:::   ||||||:::::::::
 731 PheGlnAlaSerThrProValSerGluIleGlySerLeuAsnIleGlyAs 747

 978 .................................................. 978

 747 nArgProThrSerArgLysProThrThrSerIleAlaAspLeuArgAlaI 764

 979 ...........................................GGAAAA 984
                                                   |||
 764 leProTrpValLeuAlaTrpSerGlnSerArgValMetLeuProGlyTrp 780

 985 TACGGCTTGGGCGGC..............ATCACGGCACATCCTGTCAA 1019
     ||||||   |||:::              :::::|||   |||   ::
 781 TyrGlyThrGlySerAlaPheGlnGlnTrpValAlaAlaGlyProGluSe 797

1020 GCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCA 1069
     ::::||||||      |||:::
 797 rGluSerGlnArgValGluMet........................... 804

1070 GCGACAATTTTGCCGATGCGGCATACGCCAAATACCCGTCCCCTTACCAT 1119
                          |||   ::::::|||
 805 ..........LeuHisAspLeuTyrGlnArgTrpPro.........Phe 814

1120 TCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACAT 1169
        |||:::::   ||||||:::   ||| |||   ||||||      |
 815 PheArgSerValLeuSerAsnMetAlaGln.ValLeuAlaLys......S 829

1170 CACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAA 1219
     ||            :::   |||   :::|||   :::
 829 erAspLeuGlyLeuAlaAlaArgTyrAlaGluLeuVal.........Val 842

1220 ACAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCG 1269
        :::|||   ||||||                                ||
 843 AspGluAlaLeuArgArg.............................Ar 849

1270 AATTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACCGCTGTAC. 1318
     |:::::   |||      :::      |||:::   |||   ::::::
 849 gValPheAspLysIleAlaAspGluHisArgArgThrIleAlaIleHisL 866

1319 ..............CACAAGTGAATCCTATAGATGAACCCGTCTTTAATC 1354
                     |||      :::|||      ||||||:::|||
 866 ysLeuIleThrGlyHisAspAspLeuLeuAlaAspAsnProAlaLeuAla 882

1355 CTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTGCCAGAATT 1404
        :::|||                                        ||
 883 ArgSerVal.....................................Ph 886

1405 CAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATCCCACCTAA 1454
     ||||         |||               :::   ::::::||||||:
 886 eAsnArgProTyrLeu............GluProLeuAsnHisLeuG 899

1455 AAATTACTCTCCTTCAGCACCGCTACCAAAAG.........GACCTAATA 1495
     :::::   ||||||:::   ||||||::::::         |||   :::
 899 lnValGluLeuLeuArg...ArgTyrArgSerGlyHisAspAspGluMet 914

1496 ATGGATATTTGGATAAATTTGGTAATGAATGGACTA 1531
     ::::::      |||   |||:::||||||||||||
 915 ValGlnArgGlyIleLeuLeuThrMetAsnGlyLeu 926
```

```
seq_name: SwissProt_40:SYJ1_HUMAN

seq_documentation_block:
ID   SYJ1_HUMAN     STANDARD;      PRT;  1575 AA.
AC   O43426; O43425;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Synaptojanin 1 (EC 3.1.3.56) (Synaptic inositol-1,4,5-trisphosphate 5-
DE   phosphatase 1).
GN   SYNJ1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Cerebellum;
RX   MEDLINE=98088905; PubMed=9428629;
RA   Haffner C., Takei K., Chen H., Ringstad N., Hudson A., Butler M.H.,
RA   Salcini A.E., Di Fiore P.P., De Camilli P.;
RT   "Synaptojanin 1: localization on coated endocytic intermediates in
RT   nerve terminals and interaction of its 170 kDa isoform with Eps15.";
RL   FEBS Lett. 419:175-180(1997).
CC   -!- FUNCTION: INOSITOL 5-PHOSPHATASE WHICH HAS A ROLE IN CLATHRIN-
CC       MEDIATED ENDOCYTOSIS.
CC   -!- CATALYTIC ACTIVITY: D-myo-inositol 1,4,5-trisphosphate + H(2)O =
CC       D-myo-inositol 1,4-bisphosphate + phosphate.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; A LONG ISOFORM/SYNAPTOJANIN-
CC       170 (SHOWN HERE) AND A SHORT ISOFORM/SYNAPTOJANIN-145; ARE
CC       PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: CONCENTRATED AT CLATHRIN-COATED ENDOCYTIC
CC       INTERMEDIATES IN NERVE TERMINALS. THE LONG ISOFORM IS MORE
CC       ENRICHED THAN THE SHORT ISOFORM IN DEVELOPING BRAIN AS WELL AS
CC       NON-NEURONAL CELLS. THE SHORT ISOFORM IS VERY ABUNDANT IN NERVE
CC       TERMINALS.
CC   -!- DOMAIN: BINDS TO EPS15 (A CLATHRIN COAT-ASSOCIATED PROTEIN) VIA A
CC       C-TERMINAL DOMAIN CONTAINING THREE ASN-PRO-PHE (NPF) REPEATS.
CC   -!- DOMAIN: THE C-TERMINAL PROLINE-RICH REGION MEDIATES BINDING TO A
CC       VARIETY OF SH3 DOMAIN-CONTAINING PROTEINS INCLUDING AMPHIPHYSIN,
CC       SH3P4 AND GRB2.
CC   -!- SIMILARITY: IN THE CENTRAL SECTION; BELONGS TO THE INOSITOL-1,4,5-
CC       TRISPHOSPHATE 5-PHOSPHATASE FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 SAC DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 RNA RECOGNITION MOTIF (RRM).
```

```
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF009040; AAC51922.1; -.
DR   EMBL; AF009039; AAC51921.1; -.
DR   MIM; 604297; -.
DR   InterPro; IPR000300; IPPc.
DR   InterPro; IPR000504; RRM.
DR   InterPro; IPR002013; Syja_N.
DR   Pfam; PF00783; IPPc; 1.
DR   Pfam; PF02383; Syja_N; 1.
DR   SMART; SM00128; IPPc; 1.
DR   PROSITE; PS50102; RRM; 1.
DR   PROSITE; PS50275; SAC; 1.
KW   Hydrolase; Alternative splicing; Repeat; Endocytosis; RNA-binding;
KW   Multigene family.
FT   DOMAIN      119    442       SAC.
FT   DOMAIN      500    899       CATALYTIC (POTENTIAL).
FT   DOMAIN      902    971       RNA-BINDING (RRM).
FT   DOMAIN      900   1575       PRO-RICH.
FT   DOMAIN     1033   1036       POLY-SER.
FT   DOMAIN     1108   1113       POLY-PRO.
FT   DOMAIN     1126   1129       POLY-PRO.
FT   DOMAIN     1487   1490       POLY-GLU.
FT   DOMAIN     1540   1546       POLY-PRO.
FT   DOMAIN     1396   1419       3 X 3 AA REPEATS OF N-P-F.
FT   REPEAT     1396   1398       1.
FT   REPEAT     1406   1408       2.
FT   REPEAT     1417   1419       3.
FT   VARSPLIC   1306   1311       VKTNGI -> QEQPSG (IN SHORT ISOFORM).
FT   VARSPLIC   1312   1575       MISSING (IN SHORT ISOFORM).
SQ   SEQUENCE   1575 AA;  173345 MW;  50646F6CC043B9E7 CRC64;


alignment_scores:
            Quality:  118.00              Length:    585
              Ratio:  0.470                Gaps:     34
 Percent Similarity:  42.906   Percent Identity: 24.103

alignment_block:
 US-09-303-518D-465 x SYJ1_HUMAN   ..

 Align seg 1/1  to: SYJ1_HUMAN  from: 1  to: 1575

     116 GTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGGG 165
         :::||||||:::   |||:::   :::|||:::   :::      |||
     986 IleSerIleAlaLeuProSer...SerThrSerSerThrLeuLeuGlyGl 1001

     166 GAACTTGCCGAGCGCAGCG.........GTCATATCGGATTGGGAAACAT 206
          :::          ||||||         :::   :::|||   :::   |
    1001 uAspAlaGluValAlaAlaAspPheAspMetGluGlyAspValAspAspT 1018

     207 ACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAG 256
         ||:::|||                        :::      :::|||
    1018 yrSerAlaGluValGluGluLeuLeuProGlnHisLeuGlnPro...... 1032

     257 GAAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCAT 306
         :::   |||:::   |||::::::   |||   ::::::   ::::::
    1033 SerSerSerSerGlyLeuGlyThrSerProSerSerSerProArgThrSe 1049

     307 TCCC...CCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGG 353
          |||      |||   |||:::   :::   |||:::         |||:
    1049 rProCysGlnSerProThrIleSerGluGlyProValProSerLeuProI 1066

     354 TAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACG 403
         ::   |||           :::   |||   :::
    1066 leArgProSerArgAlaProSerArgThrProGly.............. 1077

     404 AACACCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCG......GC 447
                 ||||||:::|||:::::::::         |||      ||
    1078 .........ProProSerAlaGlnSerSerProIleAspAlaGlnProAl 1091

     448 TATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGG 497
         |   |||||||||  :::::::::
    1091 aThrProLeuPro.GlnLysAsp........................... 1098

     498 CGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGAC 547
               |||       |||   :::|||   |||   |||      |||
    1099 ......ProAlaGlnProLeuGluProLysArgProProProProArgPr 1113

     548 AACGGCT......TGTCGACCGTTTCCACAATACCGGTAGTATGCTGACG 591
          :::|||         ||||||   |||      |||           :
    1113 oValAlaProProThrArgProAlaProProGlnArgProProProProS 1130

     592 CAAGGAGTAGGCGACGGATTCAAACGCGCCAC.................. 623
         ::   :::|||            ||||||
    1130 erGlyAlaArgSerProAlaProThrArgLysGluPheGlyGlyIleGly 1146

     624 ...CCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTT 670
            |||    ||||||   :::   |||           |||
    1147 AlaPro.ProSerProGlyValAlaArgArgGluMetGluAlaProLysS 1163

     671 TCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCA...... 714
               |||||||:::          |||:::   ||||||   :::
    1163 erProGlyThrThrArg......LysAspAsnIleGlyArgSerGlnPro 1177

     715 .........GGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGGTATAAG 755
                  :::   :::::::|||   |||   |||      |||   ||
    1178 SerProGlnAlaGlyLeuAlaGlyProGlyProAla......GlyTyrSe 1192

     756 CGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCG 805
         |    |||:::      :::         |||  |||         |||
    1192 r.ThrAlaArgProThrIleProProArgAlaGlyValIleSerAlaPro 1208

     806 AAAACAAGATGGCGCGCATCA.............ACGATTTGGCAGATATG 843
         :::::::::   |||||||||            |||      :::
    1209 GlnSerHisAlaArgAlaSerAlaGlyArgLeuThrProGluSerGlnSe 1225

     844 GCGCAACTCAAAGACTATG.......................CCGCAGC 869
          ::::::|||:::|||                          |||   :
    1225 rLysThrSerGluThrSerLysGlySerThrPheLeuProGluProLeuL 1242

     870 AGCCATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAG 919
         ::|||:::|||           ::::::   :::|||      |||
    1242 ysProGlnAlaAlaProProGlnSerSerLeuProProProAlaGln 1258

     920 AAG.................CCGTCAGCAATATCTTTACGGCAGTCATC 951
         :::                  |||   |||      :::   ||||||::
    1259 ArgLeuGlnGluProLeuValProValAlaAlaProMetProGlnSerGl 1275

     952 CCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCAT 1001
         :|||:::   :::||||||         :::                :::  |
    1275 yProGlnProAsnLeuGluThrProProGlnProProProArgSerArgS 1292

    1002 CACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGA 1051
         ||   |||   |||   |||   |||::::::::   ::::::
    1292 erSerHisSerLeuProSer.GluAlaSerSerGlnProGlnValLysTh 1308

    1052 AAGGGAAATC.........CGCCGTCAGCGACAATTT.............. 1079
          :::   |||         :::|||:::      :::
    1308 rAsnGlyIleSerAspGlyLysArgGluSerProLeuLysIleAspProP 1325

    1080 ....................................TGCCGATGCGGC 1091
                                             ::::::   :::
    1325 heGluAspLeuSerPheAsnLeuLeuAlaValSerLysAlaGlnLeuSer 1341

    1092 ATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATA............. 1129
         ::::::      |||||||||   |||  ||||||
    1342 ValGlnThrSerProValProThrProAspProLysArgLeuIleGlnLe 1358
```

```
1130 ..................TCCGTTCAAACTTGGAGCAGCGTTACGGCAA 1160
                        :::|||:::|||   |||||||||::::::
1358 uProSerAlaThrGlnSerAsnValAsnThrLeuSerSerValSerCysM 1375

1161 AGAAAACATCACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGA 1210
        |||   ||||||   |||:::|||   |||:::
1375 etProThrMetProProIleProAlaArgSerGlnSer............ 1387

1211 AACTGGCAAACAAACGCCACCCGAAGACCAAAGTGCCGTTTGACGGTAAA 1260
           |||   |||            |||:::  ||||||
1388 ......GlnGluAsnMetArgSerSerProAsn..ProPhe......... 1398

1261 GGGTTTCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTAAT.. 1308
                              :::         ||||||   |||
1399 .......................IleThrGlyLeuThrArgThrAsnPr 1407

1309 .............ACCGCTGTACCACAAGTGAATCCTATAGATGAACCCG 1345
                  ||||||:::|||      ||||||
1407 oPheSerAspArgThrAlaAlaProGly...AsnPro.............. 1418

1346 TCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACT 1395
       |||:::   |||::::::   ::::::|||   ||||||
1419 ..PheArgAlaLysSerGluGluSerGluAlaThrSerTrp......... 1431

1396 GCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATAT 1445
                  ::::::|||
1432 ............PheSerLys.........................Gl 1435

1446 CCCACCTAAAAATTACTCTCCTTCAGCACCGCTA...CCAAAAGGACCTA 1492
        |||   :::   ||||||         |||   |||   |||   |
1435 uGluProValThrIleSerProPheProSerLeuGlnProLeuGlyHisA 1452

1493 AT.......................AATGGATATTTGGATAAATTTGGT 1518
     ||                       :::|||:::   |||:::|||
1452 snLysSerArgAlaSerSerSerLeuAspGlyPheLysAspSerPheAsp 1468

1519 AATGAATGGACTAAAGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATG 1568
                  :::|||   |||    ||||||::::::|||   :::|
1469 .........LeuGlnGlyGlnSer.ThrLeuLysIleSerAsnProLysG 1482

1569 GGATGTTCAATTGTCTAAAACAGGAAGAGAGCAACTTGGATGGG 1612
     ||      :::      ::::::|||:::   ::::::   |||
1482 lyTrpValThrPheGluGluGluGluAspPheGlyValLysGly 1496
```

seq_name: SwissProt_40:TFE2_HUMAN

seq_documentation_block:

```
ID   TFE2_HUMAN     STANDARD;      PRT;   654 AA.
AC   P15923; P15883; Q9UPI9; Q14635; Q14636; Q14208;
DT   01-APR-1990 (Rel. 14, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Transcription factor E2-alpha (Immunoglobulin enhancer binding factor
DE   E12/E47) (Transcription factor-3) (TCF-3) (Immunoglobulin
DE   transcription factor-1) (Transcription factor ITF-1) (Kappa-E2-binding
DE   factor).
GN   TCF3 OR E2A OR ITF1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A. (ISOFORM E12).
RX   MEDLINE=90150282; PubMed=1967983;
RA   Kamps M.P., Murre C., Sun X.-H., Baltimore D.;
RT   "A new homeobox gene contributes the DNA binding domain of the
RT   t(1;19) translocation protein in pre-B ALL.";
RL   Cell 60:547-555(1990).
RN   [2]
RP   SEQUENCE FROM N.A. (ISOFORM E12).
RX   MEDLINE=90150281; PubMed=1967982;
RA   Nourse J., Mellentin J.D., Galili N., Wilkinson J., Stanbridge E.,
RA   Smith S.D., Cleary M.L.;
RT   "Chromosomal translocation t(1;19) results in synthesis of a homeobox
RT   fusion mRNA that codes for a potential chimeric transcription
RT   factor.";
RL   Cell 60:535-545(1990).
RN   [3]
RP   SEQUENCE FROM N.A. (ISOFORM E12).
RA   Lamerdin J.E., McCready P.M., Skowronski E., Viswanathan V.,
RA   Burkhart-Schultz K., Gordon L., Dias J., Ramirez M., Stilwagen S.,
RA   Phan H., Velasco N., Do L., Regala W., Terry A., Garnes J.,
RA   Danganan L., Erler A., Christensen M., Georgescu A., Avila J., Liu S.,
RA   Attix C., Andreise T., Trankheim M., Amico-Keller G., Coefield J.,
RA   Duarte S., Lucas S., Bruce R., Thomas P., Quan G., Kronmiller B.,
RA   Arellano A., Sanders C., Ow D., Nolan M., Trong S., Kobayashi A.,
RA   Olsen A.S., Carrano A.V.;
RT   "Sequence analysis of a 3.5 Mb contig in human 19p13.3 containing a
RT   serine protease gene cluster.";
RL   Submitted (JAN-1999) to the EMBL/GenBank/DDBJ databases.
RN   [4]
RP   SEQUENCE OF 214-654 FROM N.A. (ISOFORMS E12 AND E47), AND BHLH
RP   DOMAIN.
RC   TISSUE=Lymphoma;
RX   MEDLINE=89168418; PubMed=2493990;
RA   Murre C., McCaw P.S., Baltimore D.;
RT   "A new DNA binding and dimerization motif in immunoglobulin enhancer
RT   binding, daughterless, MyoD, and myc proteins.";
RL   Cell 56:777-783(1989).
RN   [5]
RP   SEQUENCE OF 69-654 FROM N.A. (ISOFORM E47).
RX   MEDLINE=90175015; PubMed=2308859;
RA   Henthorn P., McCarrick-Walmsley R., Kadesch T.;
RT   "Sequence of the cDNA encoding ITF-1, a positive-acting transcription
RT   factor.";
RL   Nucleic Acids Res. 18:677-677(1990).
RN   [6]
RP   SEQUENCE OF 511-654 FROM N.A. (ISOFORM E47).
RX   MEDLINE=92297964; PubMed=1818757;
RA   Zhang Y., Bina M.;
RT   "Sequence of a HeLa cDNA provides the DNA binding domain and carboxy
RT   terminus of HE47: a human helix-loop-helix protein related to the
RT   enhancer binding factor E47.";
RL   DNA Seq. 2:197-202(1992).
RN   [7]
RP   DISCUSSION OF SEQUENCE.
RX   MEDLINE=90140708; PubMed=2105528;
RA   Henthorn P., Kiledjian M., Kadesch T.;
RT   "Two distinct transcription factors that bind the immunoglobulin
RT   enhancer microE5/kappa 2 motif.";
RL   Science 247:467-470(1990).
RN   [8]
RP   MUTAGENESIS.
RX   MEDLINE=90280447; PubMed=2112746;
RA   Voronova A., Baltimore D.;
RT   "Mutations that disrupt DNA binding and dimer formation in the E47
RT   helix-loop-helix protein map to distinct domains.";
RL   Proc. Natl. Acad. Sci. U.S.A. 87:4722-4726(1990).
CC   -!- FUNCTION: HETERODIMERS BETWEEN TCF3 AND TISSUE-SPECIFIC BASIC
CC       HELIX-LOOP-HELIX (BHLH) PROTEINS PLAY MAJOR ROLES IN DETERMINING
CC       TISSUE-SPECIFIC CELL FATE DURING EMBRYOGENESIS, LIKE MUSCLE OR
CC       EARLY B-CELL DIFFERENTIATION. DIMERS BIND DNA ON E-BOX MOTIFS: 5'-
CC       CANNTG-3'. BINDS TO THE KAPPA-E2 SITE IN THE KAPPA IMMUNOGLOBULIN
CC       GENE ENHANCER.
CC   -!- SUBUNIT: EFFICIENT DNA BINDING REQUIRES DIMERIZATION WITH ANOTHER
CC       BHLH PROTEIN. FORMS A HETERODIMER WITH ASH1.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; E47/PAN-1 AND E12/PAN-2 (SHOWN
CC       HERE); ARE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- PTM: PHOSPHORYLATED FOLLOWING NGF STIMULATION (BY SIMILARITY).
CC   -!- DISEASE: A FORM OF PRE-B-CELL ACUTE LYMPHOBLASTIC LEUKEMIA (B-ALL)
CC       IS CHARACTERIZED BY A CHROMOSOMAL TRANSLOCATION T(1;19)(Q23;P13.3)
CC       WHICH INVOLVES PBX1 AND TCF3.
CC   -!- SIMILARITY: BELONGS TO THE BASIC HELIX-LOOP-HELIX (BHLH) FAMILY OF
CC       TRANSCRIPTION FACTORS.
CC   ---------------------------------------------------------------------
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   -----------------------------------------------------------------------
DR   EMBL; M31523; AAA61146.1; -.
DR   EMBL; M31522; AAA36764.1; ALT_SEQ.
DR   EMBL; M31222; AAA52331.1; ALT_INIT.
DR   EMBL; AC006274; AAC99797.1; -.
DR   EMBL; AC005321; AAC27373.1; -.
DR   EMBL; M24404; AAA56829.1; -.
DR   EMBL; M24405; AAA56830.1; -.
DR   EMBL; X52078; CAA36297.1; -.
DR   EMBL; M65214; AAC41693.1; -.
DR   PIR; A31492; A31492.
DR   PIR; A34734; A34734.
DR   PIR; S10099; S10099.
DR   HSSP; P10085; 1MDY.
DR   TRANSFAC; T00204; -.
DR   MIM; 147141; -.
DR   InterPro; IPR003015; HLH_Myc.
DR   InterPro; IPR001092; HLH_dim.
DR   Pfam; PF00010; HLH; 1.
DR   SMART; SM00353; HLH; 1.
DR   PROSITE; PS00038; HELIX_LOOP_HELIX; 1.
KW   Transcription regulation; DNA-binding; Nuclear protein;
KW   Proto-oncogene; Chromosomal translocation; Alternative splicing;
KW   Phosphorylation.
FT   DOMAIN      170    176       NUCLEAR LOCALIZATION SIGNAL (POTENTIAL).
FT   DOMAIN      389    425       LEUCINE-ZIPPER (POTENTIAL).
FT   DNA_BIND    547    561       BASIC DOMAIN.
FT   DOMAIN      562    605       HELIX-LOOP-HELIX MOTIF (BY SIMILARITY).
FT   SITE        483    484       BREAKPOINT FOR TRANSLOCATION TO FORM
FT                                TCF3-PBX1 ONCOGENE.
FT   VARSPLIC    530    601       PDEDEDDLLPPEQKAEREKERRVANNARERLRVRDINEAFK
FT                                ELGRMCQLHLNSEKPQTKLLILHQAVSVILN -> STDEVL
FT                                SLEEKDLRDRERRMANNARERVRVRDINEAFRELGRMCQMH
FT                                LKSDKAQTKLLILQQAVQVILG (IN ISOFORM E47).
FT   MUTAGEN     550    551       RR->GG: NO DNA-BINDING.
FT   MUTAGEN     551    551       R->K: NO DNA-BINDING.
FT   MUTAGEN     561    561       R->K: NO DNA-BINDING.
FT   MUTAGEN     561    563       RVR->GVG: NO DNA-BINDING.
FT   MUTAGEN     563    563       R->K: NO DNA-BINDING.
FT   MUTAGEN     588    588       K->A: NO DNA-BINDING AND NO
FT                                DIMERIZATION.
FT   MUTAGEN     591    592       IL->DE: NO DNA-BINDING AND NO
FT                                DIMERIZATION.
FT   MUTAGEN     595    595       A->D: NO CHANGE IN DNA-BINDING OR
FT                                DIMERIZATION.
FT   CONFLICT     69     99       FDPSRTFSEGTHFTESHSSLSSSTFLGPGLG -> GGGECL
FT                                AWCGPSAVHRCADVGLGMVSARTAP (IN REF. 4).
FT   CONFLICT    214    216       FYV -> EFR (IN REF. 4; AAA56830).
FT   CONFLICT    390    390       MISSING (IN REF. 4; AAA56830).
FT   CONFLICT    552    552       V -> M (IN REF. 4).
FT   CONFLICT    560    560       L -> V (IN REF. 4).
FT   CONFLICT    570    570       K -> R (IN REF. 4).
FT   CONFLICT    578    578       L -> M (IN REF. 4).
FT   CONFLICT    581    585       NSEKP -> KSDKA (IN REF. 4).
FT   CONFLICT    593    593       H -> Q (IN REF. 4).
FT   CONFLICT    597    597       S -> Q (IN REF. 4).
FT   CONFLICT    601    601       N -> G (IN REF. 4).
SQ   SEQUENCE   654 AA;  67600 MW;  52F5E3DE1890AE13 CRC64;
```

```
alignment_scores:
            Quality:  117.50            Length:    600
              Ratio:  0.461              Gaps:      33
 Percent Similarity:  42.500   Percent Identity: 21.333

alignment_block:
```

```
US-09-303-518D-465 x TFE2_HUMAN   ..

Align seg 1/1  to: TFE2_HUMAN  from: 1  to: 654

     12 CCGCAAAATATCCCTTATTCTGTCCATACTGGCAGTGTGCCTGCCGATGC 61
        |||   ::::::|||   ||||||:::      |||         |||
    177 ProProGlyLeuProSerSerValTyrProProSerSerGlyGluAsp.. 192

     62 ATGCACACGCCTCAGATTTGGCAAA...CGATTCTTTTATCCGGCAGGTT 108
                      :::|||:::           ||||||:::::::
    193 .............TyrGlyArgAspAlaThrAlaTyrProSerAlaL 204

    109 CTCGACCGTCAGCATTTCGAACCC.................GACGGGAA 140
        ::   ||||||:::       |||                ||||||::
    204 ysThrProSerSer.ThrTyrProAlaProPheTyrValAlaAspGlySe 220

    141 ATACCAC............CTATTCGGCAGCAGGGGGGAA.......... 168
        :   |||           |||::::::      |||:::
    220 rLeuHisProSerAlaGluLeuTrpSerProProGlyGlnAlaGlyPheG 237

    169 ................................CTTGCCGAGCGCAGC 183
                                        |||         |||
    237 lyProMetLeuGlyGlyGlySerSerProLeuProLeuProProGlySer 253

    184 GGTCATATCGGA.....................TTGGGAAACATACAAAG 212
        |||   :::|||                    :::|||:::::::::::
    254 GlyProValGlySerSerGlySerSerSerThrPheGlyGlyLeuHisGl 270

    213 CCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATA 262
        :|||:::                                    ::::
    270 nHisGlu.....................................ArgM 274

    263 TCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCC 312
        ::||||||   :::            ||||||   ||||||:::::
    274 etGlyTyrGlnLeu............HisGlyAlaGluValAsnGlyGly 286

    313 TTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGT 362
        :::   :::   :::|||   |||::::::        |||:::
    287 LeuProSerAlaSerSerPheSerSerAla......ProGlyAlaThrTy 301

    363 TGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACCATC 412
            |||   |||                          :::|||   |
    301 rGlyGlyValSer.........................SerHisThrP 309

    413 CC............GCCGACGGCTATGACGGGCCACAGGGCGGCGGCTAT 450
        ||            ||||||:::     |||   :::|||
    309 roProValSerGlyAlaAspSerLeuLeuGlySerArgGly......... 322

    451 CCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGT 500

    322 ..................................................... 322

    501 TGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAAC 550
                                  |||         |||:::|||:::
    323 .......................ThrThrAlaGlySerSerGlyAspA 331

    551 GGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTA 600
          |||
    331 laLeu.................................................. 332

    601 GGCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATC 650
            |||         |||:::   |||||||||      |||:::||
    333 ......GlyLysAlaLeuAlaSerIleTyrSerPro......AspHisSe 345

    651 GGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCGTCAAAAACA 700
        |:::|||      ::::::      :::|||          |||   :::
    345 rSerAsnAsnPheSerSerSerProSerThrPro......ValGlySerP 360

    701 TCATCGGCGCGGCAGGAGAAATT............GTCGGCGCAGGCGAT 738
            |||   ||||||                  :::||||||
    360 roGlnGlyLeuAlaGlyThrSerGlnTrpProArgAlaGlyAlaProGly 376

    739 GCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTT 788
```

```
     |||::::::         :::|||:::            :::||||||||
 377 AlaLeuSerProSerTyrAspGlyGly...........LeuHisGlyLe 389

 789 GGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAG 838
     |                        :::::::::|||:::|||      |
 389 u.....................GlnSerLysIleGluAspHisLeuA 398

 839 ATATGGCG......CAACTCAAAGACTATGCCGCAGCAGCCATCCGCGAT 882
     ||   |||         |||:::::::::|||:::::::      |||
 398 spGluAlaIleHisValLeuArgSerHisAlaValGlyThrAlaGlyAsp 414

 883 TGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAA 932
         :::      |||           |||      |||::::::::
 415 MetHisThrLeuLeuPro...........GlyHisGlyAlaLeuAlaSe 427

 933 TATCTTTACGGCAGTCATCCCCGTCAAAGGG......ATTGGAGCTGTTC 976
     :   ||||||:::   :::   :::   |||        |||   |||
 427 rGlyPheThrGlyProMetSerLeuGlyGlyArgHisAlaGlyLeuValG 444

 977 GGGGAAAATAC.........GGCTTGGGCGGCATCACGGCA......... 1008
       |||::::::         ||||||:::|||   |||:::
 444 lyGlySerHisProGluAspGlyLeuAlaGlySerThrSerLeuMetHis 460

1009 ...CATCCTGTCAAGCGGTCGCAGATGGGCGAGATC...........GC 1043
        |||   :::      ||||||   |||   :::
 461 AsnHisAlaAlaLeuProSerGlnProGlyThrLeuProAspLeuSerAr 477

1044 ATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCAT 1093
        |||   :::   |||:::::::::          |||   ||||||
 477 gProProAspSerTyrSerGlyLeuGlyArgAlaGlyAlaThrAlaAlaA 494

1094 ACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTG 1143
       ::::::                                    :::
 494 laSerGluIle..............................LysArg 499

1144 GAGCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCCGCCGTC 1193
     |||::::::      :::|||||||   :::::::::          ::
 500 GluGluLysGluAspGluGluAsnThrSerAlaAlaAspHisSerGluGl 516

1194 AAACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAA. 1242
     ::::   |||::::::|||      :::   |||              :::
 516 uGluLysLysGluLeuLysAlaProArgAlaArgThrSerProAspGluA 533

1243 ..............GTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAA 1278
                  :::|||   :::   |||:::      :::   |||
 533 spGluAspAspLeuLeuProProGluGlnLysAlaGluArgGluLysGlu 549

1279 AAAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAA 1328
     :::   |||      ::::::|||   :::      |||   ::::::||
 550 ArgArgValAlaAsnAsnAlaArgGluArgLeuArgValArgAspIleAs 566

1329 TCCTATAGATGAACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTC 1378
     |              :::|||:::      |||   :::   :::   |
 566 n............GluAlaPheLysGluLeuGlyArgMetCysGlnLeuH 579

1379 ATTCTTGGTCTATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAA 1428
     || |||                     |||::::::::::   |||   ||
 579 is.Leu.....................AsnSerGluLysProGlnThrLy 588

1429 GGTAGAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCT 1478
     |:::
 588 sLeu.......................................... 589

1479 ACCAAAAGGACCTAATAATGGATATTTGGATAAATTTGGTAATGAATGGA 1528
                |||||||:::      ::::::::::::::|||
 590 ...........LeuIleLeuHisGlnAlaValSerValIleLeuAsn... 601

1529 CTAAAGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATGGGATGTTCAA 1578
           ::::::|||::::::::::   :::||||||||||
 602 ......LeuGluGlnGlnValArgGluArgAsnLeuAsnProLysAlaAl 616

1579 TTGTCTAAAA...CAGGAAGAGAGCAACTTGGATGGGCTAGTAGGG 1621
      |||||||||   :::||||||::::::::::::|||:::||||||
```

```
     616 aCysLeuLysArgArgGluGluGluLysValSerGlyValValGly 631

seq_name: SwissProt_40:KIN4_YEAST

seq_documentation_block:
ID   KIN4_YEAST      STANDARD;      PRT;   800 AA.
AC   Q01919;
DT   01-FEB-1995 (Rel. 31, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Serine/threonine-protein kinase KIN4 (EC 2.7.1.-).
GN   KIN4 OR KIN31 OR KIN3 OR YOR233W OR O5220.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=93220392; PubMed=8465601;
RA   Kambouris N.G., Burke D.J., Creutz C.E.;
RT   "Cloning and genetic analysis of the gene encoding a new protein
RT   kinase in Saccharomyces cerevisiae.";
RL   Yeast 9:141-150(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / FY1679;
RX   MEDLINE=97127829; PubMed=8972580;
RA   Boyer J., Michaux G., Fairhead C., Gaillon L., Dujon B.;
RT   "Sequence and analysis of a 26.9 kb fragment from chromosome XV of
RT   the yeast Saccharomyces cerevisiae.";
RL   Yeast 12:1575-1586(1996).
CC   -!- FUNCTION: THIS PROTEIN IS PROBABLY A SERINE/THREONINE PROTEIN
CC       KINASE.
CC   -!- SIMILARITY: BELONGS TO THE SER/THR FAMILY OF PROTEIN KINASES.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67916; CAA48115.1; -.
DR   EMBL; Z75141; CAA99453.1; -.
DR   PIR; S29344; S29344.
DR   HSSP; Q63450; 1A06.
DR   SGD; S0005759; KIN4.
DR   InterPro; IPR000719; Euk_pkinase.
DR   InterPro; IPR002290; Ser_thr_pkinase.
DR   Pfam; PF00069; pkinase; 1.
DR   SMART; SM00220; S_TKc; 1.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; 1.
DR   PROSITE; PS00108; PROTEIN_KINASE_ST; 1.
DR   PROSITE; PS50011; PROTEIN_KINASE_DOM; 1.
KW   Transferase; Serine/threonine-protein kinase; ATP-binding.
FT   DOMAIN       46    313       PROTEIN KINASE.
FT   NP_BIND      52     60       ATP (BY SIMILARITY).
FT   BINDING      80     80       ATP (BY SIMILARITY).
FT   ACT_SITE    175    175       BY SIMILARITY.
SQ   SEQUENCE   800 AA;  90087 MW;  655BBB5EBD8ACF65 CRC64;


alignment_scores:
            Quality:  117.50              Length:    363
              Ratio:  0.656                 Gaps:     19
 Percent Similarity:  49.311   Percent Identity: 24.242

alignment_block:
 US-09-303-518D-465 x KIN4_YEAST   ..

 Align seg 1/1  to: KIN4_YEAST  from: 1  to: 800
```

```
 478 TACAGCTACGACATAAAAGGCGTTGCCCAA.................AA 509
     |||:::   |||   ||||||   :::|||                  ::
 422 TyrThrArgAspSerLysGlyAsnGlyGlnThrGlnIleGluGlnValSe 438

 510 TATCCGCCTCAACCTGACCGACAACCGCAGCACC.............. 543
     :   |||   :::         |||:::   |||
 438 rAlaArgHisSerSerArgGlyAsnLysHisThrSerValAlaGlyLeuV 455

 544 ..........GGACAACGGCTTGTCGACCGTTTCCACAATACC...GGT 579
             |||:::      :::   |||   :::|||:::   :::
 455 alThrIleProGlySerProThrThrAlaArgThrArgAsnAlaProSer 471

 580 AGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAA.............. 615
     |||   ||||||:::   |||   |||:::   :::
 472 SerLysLeuThrGluHisValLysAspSerSerGlnThrSerPheThrGl 488

 616 ...............................................C 616
                                                   |
 488 nGluGluPheHisArgIleGlyAsnTyrHisValProArgSerArgProA 505

 617 GCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAA 666
     ||   |||   |||   |||   |||:::|||      |||:::|||:::
 505 rgProThrSerTyrTyrProGlyLeuSerArg......AsnThrAlaAsp 519

 667 GCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGG 716
         |||:::   |||||||||      |||   :::|||:::   ||
 520 ......AsnSerLeuAlaAspIleProValAsnLysLeuGlySerAsnGl 534

 717 AGAAATTGTCGGCGCAGGCGATGCCGTG......CAGGGTATAAGCGAAG 760
     |:::::::::   |||   |||   |||      :::::::|||   :::
 534 yArgLeuThrAspAlaLysAspProValProLeuAsnAlaIleHisAspT 551

 761 GCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAAC 810
       ::::::   ::::::   ::::::      :::|||   :::|||:::
 551 hrAsnLysAlaThrIleSerAsnAsnSerIleMetLeuLeuSerGluGly 567

 811 AAGATGGCGCGC.........................ATCAACGATTT 833
         |||:::                           |||:::|||||
 568 ProAlaAlaLysThrSerProValAspTyrHisTyrAlaIleGlyAspLe 584

 834 G............GCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAG 871
     |                  ::::::::::::   |||
 584 uAsnHisGlyAspLysProIleThrGluValIleAspLysIleAsnLysA 601

 872 CCATCCGCGATTGGGCAGTCCAAAAC......CCCAATGCCGCACAAGGC 915
       :::         |||::::::|||      |||:::   :::
 601 spLeuThrHisLysAlaAlaGluAsnGlyPheProArgGluSerIleAsp 617

 916 ATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGAT 965
        |||::::::|||:::|||:::::::::      |||:::::::::
 618 ProGluSerThrSerThrIleLeuValThrLysGluProThrAsnSer.. 633

 966 TGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTG 1015
                                          |||   :::
 634 ....................................ThrAspGluAspH 638

1016 TCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCC... 1062
       :::||||||:::   :::::::::      ::::::||||||
 638 isValGluSerGlnLeuGluAsnValGlyHisSerSerAsnLysSerAsp 654

1063 GCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCAAA........... 1101
     |||   ||||||:::   :::         |||   |||
 655 AlaSerSerAspLysAspSerLysLysIleTyrGluLysLysArgPheSe 671

1102 .TACCCGTCCCCTTACCATTCCCGAAAT...ATCCGTTCAAACTTGGAGC 1147
      :::   |||   |||   |||   |||      ||||||::::::|||:
 671 rPheMetSerLeuTyrSerSerLeuAsnGlySerArgSerThrValGluS 688

1148 AGCGTTACGGCAAAGAAAACATC......ACCTCCTCAACCGTGCCG... 1188
     ::|||   :::|||   |||         :::||||||      |||
 688 erArgThrSerLysGlyAsnAlaProProValSerSerArgAsnProSer 704

1189 ...CCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAA......CGCCA 1229

          ||||||   :::|||:::|||:::::::::::::      |||::
 705 GlyGlnSerAsnArgSerAsnIleLysIleThrGlnGlnGlnProArgAs 721

1230 CCCGAAGACCAAAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAA 1279
     :   :::   :::|||                        ||||||   :::|
 721 nLeuSerAspArgVal.....................ProAsnProAspL 731

1280 AAGACGTAAAATACGATACGAGAATTAATACCGCTGTACCACAAGTGAAT 1329
     ||   ::::::   |||:::||||||:::      :::|||:::
 731 ysLysIleAsn...AspAsnArgIleArgAspAsnAlaProSerTyrAla 746

1330 CCTATAGATGAACCCGTCTTTAATCCTAAAGGTTCTGTC 1368
          :::::::|||      :::   ::::::||||||
 747 GluSerGluAsnProGlyArgSerValArgAlaSerVal 759
```

```
seq_name: SwissProt_40:CHI2_COCIM

seq_documentation_block:
ID   CHI2_COCIM     STANDARD;      PRT;   860 AA.
AC   P54197;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   01-OCT-1996 (Rel. 34, Last annotation update)
DE   Endochitinase 2 precursor (EC 3.2.1.14).
GN   CTS2.
OS   Coccidioides immitis.
OC   Eukaryota; Fungi; Ascomycota; Pezizomycotina; Eurotiomycetes;
OC   Onygenales; mitosporic Onygenales; Coccidioides.
OX   NCBI_TaxID=5501;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=C735;
RX   MEDLINE=96144270; PubMed=8566773;
RA   Pishko E.J., Kirkland T.N., Cole G.T.;
RT   "Isolation and characterization of two chitinase-encoding genes
RT   (cts1, cts2) from the fungus Coccidioides immitis.";
RL   Gene 167:173-177(1995).
CC   -!- FUNCTION: MAY BE ASSOCIATED WITH ENDOSPORULATION.
CC   -!- CATALYTIC ACTIVITY: Hydrolysis of the 1,4-beta-linkages of N-
CC       acetyl-D-glucosamine polymers of chitin.
CC   -!- SIMILARITY: BELONGS TO CHITINASE CLASS II (FAMILY 18 OF GLYCOSYL
CC       HYDROLASES).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L41662; AAA92642.1; -.
DR   HSSP; P23472; 2HVM.
DR   InterPro; IPR001579; Chitinase_2.
DR   Pfam; PF00192; chitinase_2; 1.
DR   PROSITE; PS01095; CHITINASE_18; 1.
KW   Hydrolase; Glycosidase; Chitin degradation; Chitin-binding; Signal;
KW   Glycoprotein.
FT   SIGNAL        1     22       POTENTIAL.
FT   CHAIN        23    860       ENDOCHITINASE 2.
FT   CARBOHYD     90     90       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   860 AA;  91395 MW;  5E34B54FAA663F3C CRC64;


alignment_scores:
            Quality:  117.50              Length:    382
              Ratio:  0.646                 Gaps:     16
 Percent Similarity:  47.644   Percent Identity: 21.466

alignment_block:
 US-09-303-518D-465 x CHI2_COCIM   ..

 Align seg 1/1  to: CHI2_COCIM  from: 1  to: 860
```

```
 98 TCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCAC 147
    |||::::::   :::|||::::::   ::::::   :::::::::|||||
362 SerSerGlnSerThrThrMetGluThrLysThrLeuSerAlaSerThrTh 378

148 CTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATT 197
    |   |||:::  :::         ||||||:::::::::  :::
378 rProSerSerProSerThrValSerProSerSerThrMetGlnThrThrS 395

198 GGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGG 247
                :::|||:::  :::|||     |||:::::::
395 erThrGlySerThrSerIleGluThrValThrThrArgSerGlnGluPro 411

248 CCATTAAAGGAAATATCGGCT...ACATTGTCCGCTTTTCCGATCACGGG 294
    |||         ||||||:::  :::  |||:::  |||:::|||
412 ProSerThrThrIleSerThrArgSerAlaSerThrGluProValThrTh 428

295 CACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGA 344
      ::::::  ||||||||||||||||:::
428 rArgSerGlnGluProProSerThrThrIleSer..............  439

345 TGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGG 394
                    |||  ||||||  |||  ::::::
440 ..................ThrArgSerAlaSerThrGluThrValThr 449

395 ACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCCACAGGGCGGC 444
    |||  ::::::     |||  |||::::::::::
450 ThrArgSerGlnGluProProSerThrThrIleSer.............  461

445 GGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCTACGACATAAA 494

461 ................................................  461

495 AGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGCACCG 544
                    ||||||||||||     :::|||::::::
462 ..........ThrTrpSerAlaSerThrGluThrSerThrSerSerGln 474

545 GACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGACGCAA 594
    |||:::     :::|||:::||||||
475 AspSerProSerThrThrIleSerThr.....................  483

595 GGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGCC..........  634
                  ::::::||||||     |||:::
484 ..............LysSerAlaProThrGlyThrValThrThrArgS 495

635 ..........CCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCA 673
            ||||||  |||     :::     |||:::  :::
495 erGlnAspLeuProSerThrThrIleSerThrArgSerProGluThrGlu 511

674 ACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATT 723
    |||          ::::::|||:::::::::::::  :::     ||
512 ThrGlu...ThrAlaThrThrLysSerGlnGlySerProSerIleThrLe 527

724 GTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGC 773
    ||||:::::::::                 :::::::|||:::|||:::
527 uSerThrArgSer.................SerSerAlaGluThrVal. 537

774 TGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCA 823
                                  :::||||||  :::
538 ............................SerThrArgSerGlnHis 543

824 TCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCC 873
    |||:::         |||      ::::::|||:::  |||  :::
544 SerSerSerThrThrIle......SerThrLysSerAlaProThrGluTh 558

874 ATCCGCGATTGGGCAGTCCAAAACCCCAATGCCG.........CACAAGG 914
     ::::::  ::::::  :::|||  ||||||        ::::::|
558 rGlyThrThrSerGluHisSerThrSerMetProValSerThrArgSerA 575

915 CATAGAAGCCGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGA 964
    ||      ::::::|||:::     |||:::|||  |||::::::
575 laSerThrGluThrValIleThrArgSerGlnAsnSerAspSerGlnSer 591

965 TTGGAGCTGTTCGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCT 1014
    :::  :::        :::|||        :::::::::  |||
592 MetThrVal.....SerThrArgSerProSerThrGluSerIleTh 605

1015 GTCAAGCGGTCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGC 1064
      :::                           |||:::  :::  |
605 rThr.........................ArgSerGlnGlySerP 612

1065 CGTCAGCGACAATT...........TTGCCGATG...........CGG 1090
     |||||  |||:::           :::|||:::
612 roSerGluThrPheSerThrLysSerValProValAspThrIleSerThr 628

1091 CATACGCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAAC 1140
     :::  |||:::     |||  ||||||     |||:::
629 GluLeuProSerGlnThrProThrThrIle.....IleThrGlyThrPr 643

1141 TTGGAGCAGCGTTACGGCAAAGAAAACATCA...CCTCCTCAACCG 1183
         :::  |||:::|||     |||:::  ||||||:::|||
643 oSerAspProValSerAlaProThrThrThrValProProAsnPro 658
```

seq_name: SwissProt_40:ITF2_MOUSE

seq_documentation_block:

ID ITF2_MOUSE STANDARD; PRT; 670 AA.
AC Q60722; Q64072; Q60721; Q62211;
DT 15-DEC-1998 (Rel. 37, Created)
DT 15-DEC-1998 (Rel. 37, Last sequence update)
DT 16-OCT-2001 (Rel. 40, Last annotation update)
DE Transcription factor 4 (Immunoglobulin transcription factor 2) (ITF-2)
DE (MITF-2) (SL3-3 enhancer factor 2) (SEF-2) (Class A helix-loop-helix
DE transcription factor ME2).
GN TCF4 OR ITF2 OR SEF2.
OS Mus musculus (Mouse).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX NCBI_TaxID=10090;
RN [1]
RP SEQUENCE FROM N.A. (ISOFORM MITF-2B).
RC TISSUE=Brain;
RX MEDLINE=95075243; PubMed=7984047;
RA Soosaar A., Chiaramello A., Zuber M.X., Neuman T.;
RT "Expression of basic-helix-loop-helix transcription factor ME2 during
RT brain development and in the regions of neuronal plasticity in the
RT adult brain.";
RL Brain Res. Mol. Brain Res. 25:176-180(1994).
RN [2]
RP SEQUENCE FROM N.A. (ISOFORMS MITF-2A; MITF-2B AND MITF-2B-DELTA).
RX MEDLINE=96216452; PubMed=8631961;
RA Skerjanc I.S., Truong J., Filion P., McBurney M.W.;
RT "A splice variant of the ITF-2 transcript encodes a transcription
RT factor that inhibits MyoD activity.";
RL J. Biol. Chem. 271:3555-3561(1996).
RN [3]
RP SEQUENCE FROM N.A. (ISOFORM MITF-2B).
RC STRAIN=BALB/C; TISSUE=Brain;
RX MEDLINE=97133297; PubMed=8978694;
RA Pscherer A., Doerflinger U., Kirfel J., Gawlas K., Rueschoff J.,
RA Buettner R., Schuele R.;
RT "The helix-loop-helix transcription factor SEF-2 regulates the
RT activity of a novel initiator element in the promoter of the human
RT somatostatin receptor II gene.";
RL EMBO J. 15:6680-6690(1996).
CC -!- FUNCTION: TRANSCRIPTION FACTOR THAT BINDS TO THE IMMUNOGLOBULIN
CC ENCHANCER MU-E5/KE5-MOTIF. MITF-2B INHIBITS MYOD ACTIVATION OF THE
CC CARDIAC ALPHA-ACTIN PROMOTER. BINDS TO THE E-BOX PRESENT IN THE
CC SOMATOSTATIN RECEPTOR 2 INITIATOR ELEMENT (SSTR2-INR) TO ACTIVATE
CC TRANSCRIPTION. MAY HAVE A REGULATORY FUNCTION IN DEVELOPMENTAL
CC PROCESSES AS WELL AS DURING NEURONAL PLASTICITY.
CC -!- SUBUNIT: EFFICIENT DNA BINDING REQUIRES DIMERIZATION WITH ANOTHER
CC BHLH PROTEIN. MITF-2B SEEMS TO FORM INACTIVE HETERODIMERS WITH
CC MYOD.
CC -!- SUBCELLULAR LOCATION: Nuclear (Probable).
CC -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; MITF-2A AND MITF-2B (SHOWN

```
CC       HERE); DIFFERING IN THEIR N-TERMINI, ARE PRODUCED BY ALTERNATIVE
CC       SPLICING. IN ADDITION, A MINI-EXON CODING FOR FOUR AMINO ACIDS MAY
CC       OR MAY NOT BE INCLUDED.
CC   -!- TISSUE SPECIFICITY: EXPRESSED IN THE CEREBRAL CORTEX, PURKINJE AND
CC       GRANULE CELL LAYERS OF THE CEREBELLUM, OLFACTORY NEUROEPITHELIUM,
CC       PYRAMIDAL CELLS OF HIPPOCAMPAL LAYERS CA1-CA4, AND IN THE GRANULAR
CC       CELLS OF THE DENTATE GYRUS.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED IN PROLIFERATIVE ZONES DURING
CC       DEVELOPMENT AND IN THE ADULT IN AREAS OF NEURONAL PLASTICITY. AT
CC       EMBRYONIC DAY 12 (E12), EXPRESSION IS LOCALIZED IN THE CORTEX,
CC       CEREBELLUM, PONS, MEDULLA AND SPINAL CORD. FROM E18 TO ADULTHOOD,
CC       HIGH LEVELS OF EXPRESSION ARE FOUND IN THE PYRAMIDAL CELLS OF
CC       HIPPOCAMPAL LAYERS CA1-CA4, AND IN THE GRANULAR CELLS OF THE
CC       DENTATE GYRUS. AT POSTNATAL DAY 7, EXPRESSION IS HIGH IN THE
CC       VISUAL CORTEX AND IN THE SUBEPENDYMAL REGION EXTENDING FROM THE
CC       ANTERIOR LATERAL VENTRICLE INTO THE OLFACTORY BULB.
CC   -!- SIMILARITY: BELONGS TO THE BASIC HELIX-LOOP-HELIX (BHLH) FAMILY OF
CC       TRANSCRIPTION FACTORS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; S75870; AAB32662.1; -.
DR   EMBL; U16321; AAC52414.1; -.
DR   EMBL; U16322; AAC52415.1; -.
DR   EMBL; X91753; CAA62868.1; -.
DR   MGD; MGI:98506; Tcf4.
DR   InterPro; IPR003015; HLH_Myc.
DR   InterPro; IPR001092; HLH_dim.
DR   Pfam; PF00010; HLH; 1.
DR   SMART; SM00353; HLH; 1.
DR   PROSITE; PS00038; HELIX_LOOP_HELIX; 1.
KW   Transcription regulation; DNA-binding; Activator; Nuclear protein;
KW   Alternative splicing; Developmental protein.
FT   DOMAIN        1     83       ESSENTIAL FOR MYOD INHIBITION.
FT   DOMAIN      227    230       POLY-SER.
FT   DOMAIN      378    399       LEUCINE-ZIPPER (POTENTIAL).
FT   DNA_BIND    567    579       BASIC DOMAIN.
FT   DOMAIN      580    621       HELIX-LOOP-HELIX MOTIF (BY SIMILARITY).
FT   DOMAIN      622    645       CLASS A SPECIFIC DOMAIN.
FT   VARSPLIC      1    159       MISSING (IN ISOFORM MITF-2A).
FT   VARSPLIC    160    182       LHSSAMEVQTKKVRKVPPGLPSS -> MYCAYTIPGMGGNS
FT                                LMYYYNGKA (IN ISOFORM MITF-2A).
FT   VARSPLIC    545    548       MISSING (IN ISOFORM MINI-EXON LESS).
FT   CONFLICT    159    159       P -> S (IN REF. 3).
FT   CONFLICT    180    180       P -> A (IN REF. 3).
FT   CONFLICT    238    238       G -> A (IN REF. 1).
SQ   SEQUENCE   670 AA;  71624 MW;  F21CF79FDC0BF09E CRC64;


alignment_scores:
            Quality:  117.00            Length:    553
              Ratio:  0.470              Gaps:     32
 Percent Similarity:  45.027   Percent Identity: 22.242

alignment_block:
 US-09-303-518D-465 x ITF2_MOUSE   ..

 Align seg 1/1  to: ITF2_MOUSE  from: 1  to: 670

    137 GGAAATACCACCTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGT 186
        |||      :::   :::   |||:::::::::::::||||||:::
    199 GlyTyrProSerSerLysProAlaAlaSerThrPheProSerSerPhePh 215

    187 CATATCGGATTGGGAAACAT.................ACAAAGCCATCA 218
         ::::::||| |||   |||                  ::::::   :::
    215 eMetGlnAsp.GlyHisHisSerSerAspProTrpSerSerSerSerGly 231

    219 GTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCT 268
        ::::::|||||   :::      |||||| :::   ||||||   ::::
    232 MetAsnGlnProGlyTyr......GlyGly.MetLeuGlyAsnSerSerH 246

    269 ACATTGTCCGCTTTTCCGAT............CACGGGCACGAAGTCCAT 306
        ::|||   :::   |||:::            |||   ||||||
    246 isIleProGlnSerSerSerTyrCysSerLeuHisProHisGluArgLeu 262

    307 TCCCCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAG 356
        |||            :::   |||||||||:::::|||      :::||
    263 Ser............TyrProSerHisSerSerAlaAspIleAsnSerSe 275

    357 T......CCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGAT 400
        |      |||::::::   |||              |||   :::|||
    275 rLeuProProMetSerThrPhe..............HisArgSerGlyT 287

    401 ACGAACACCAT.....................CCCGCCGACGGCTATGAC 429
          :::|||:::                     ||||||:::|||   |||
    287 hrAsnHisTyrSerThrSerSerCysThrProProAlaAsnGlyThrAsp 303

    430 GGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATA 479

    303 ................................................. 303

    480 CAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCG 529
                                                  ::::::
    304 .....................................SerIleMetA 307

    530 ACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGT 579
          ||||||:::||||||                        :::|||
    307 laAsnArgGlyThrGlyAla.....................AlaGly 315

    580 AGTATGCTGACGCAAGGAGTAGGCGAC......GGATTCAAACGCGCCAC 623
        |||      |||        ||||||      |||         |||::
    316 SerSerGlnThr.........GlyAspAlaLeuGlyLysAlaLeuAlaSe 329

    624 CCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCA 673
        :   |||||||||      |||:::::::::|||:::   ::::::   |
    329 rIleTyrSerPro......AspHisThrAsnAsnSerPheSerSer...A 343

    674 ACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATT 723
        ||   ::::::   :::   :::            :::||||||
    343 snProSerThrProValGlySerProProSerLeuSerAlaGlyThrAla 359

    724 GTCGGCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGC 773
        |||   :::   ::::::   |||:::   |||:::   :::      ::
    360 ValTrpSerArgAsnGlyGlyGlnAlaSerSerSerProAsnTyrGluGl 376

    774 TGTTATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCA 823
        :   :::|||:::|||                |||::::::::::   |||:
    376 yProLeuHisSerLeu...GlnSerArgIleGluAspArgLeuGluArgL 392

    824 TCAACGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCC 873
        :::::|||                                |||      :::
    392 euAspAsp............................AlaIleHisVal 398

    874 ATCCGCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGC 923
        :::|||:::   ||||||         :::|||:::   |||
    399 LeuArgAsnHisAlaValGlyProSerThrAlaValProGly........ 412

    924 CGTCAGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTG 973
                                            |||   |||   :
    413 .....................................GlyHisGlyAspM 417

    974 TTCGGGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGG 1023
        :::::|||      |||                      |||      :::
    417 etHisGlyIleMetGly....................ProSerHisAsn 426

   1024 TCGCAGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGA 1073
        :::   ||||||:::::::::         |||   |||   :::|||
    427 GlyAlaMetGlySerLeuGlySerGlyTyrGlyThrSerLeuLeuSer.. 442

   1074 CAATTTTGCCGATGCGGCATACG........................... 1096
```

```
          |||::: ||||||:::
  443 .......AlaAsn.ArgHisSerLeuMetValGlyAlaHisArgGluAsp 456

 1097 ..............................CCAAATACCCGTCCCCT 1113
                                    ||||||      |||
  457 GlyValAlaLeuArgGlySerHisSerLeuLeuProAsnGlnValPro.. 472

 1114 TACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGA 1163
       :::|||::::::   ||||||         |||:::|||:::
  473 ....ValProGlnLeuProValGln.........SerAlaThrSerProA 485

 1164 AAACATCACCTCCTCAACC.............GTGCCGCCGTCA...AAC 1197
       :::||||||||||              :::|||||:::   :::
  485 spLeuAsnProProGlnAspProTyrArgGlyMetProProGlyLeuGln 501

 1198 GGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGTGCC 1247
      |||:::::::|||:::   ::::::                :::::::::
  502 GlyGlnSerValSerSerGlySer..............SerGluIleLy 513

 1248 GTTTGACGGTAAAGGGTTTCCGAATTTTGAA.................. 1278
          |||   :::|||      |||::::::
  513 sSerAspAspGluGlyAspGluAsnLeuGlnAspThrLysSerSerGluA 530

 1279 ......................AAAGACGTAAAATACGATACGAGA... 1302
                             ||||||:::|||      ||||||
  530 spLysLysLeuAspAspAspLysLysAspIleLysSerIleThrArgSer 546

 1303 ......ATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCGT 1346
            |||:::          :::::::::|||   ::::::
  547 ArgSerSerAsnAsnAspAspGluAspLeuThrProGluGlnLysAlaGl 563

 1347 CTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACTG 1396
       :::   :::      :::::::::              :::   :
  563 uArgGluLysGluArgArgMetAlaAsnAsnAlaArgGluArgLeuArgV 580

 1397 CCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATATC 1446
      ::|||              :::   :::   ||||||:::      :::
  580 alArgAspIleAsnGluAlaPheLysGluLeuGlyArgMetValGlnLeu 596

 1447 CCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATAA 1496
       |||||||||           |||           |||   ||||||:
  597 .HisLeuLysSerAspLysProGln.......:.ThrLysLeuLeuIleL 610

 1497 TGGATATTTGGATAAATTTGGTAATGAATGGACTAAAGGTCCATCAAGAA 1546
      ::          :::   ::::::         |||:::::::::|||:::
  610 euHisGlnAlaValAlaValIle.........LeuSerLeuGluGlnGln 623

 1547 CTAAAGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACAGGAAGA 1596
      ::::::   :::||||||||||            |||||||||::::::||
  624 ValArgGluArgAsnLeuAsnProLysAlaAlaCysLeuLysArgArgGl 640

 1597 G 1597
      |
  640 u 640
```

```
seq_name: SwissProt_40:TEGU_HSVEB

seq_documentation_block:
ID   TEGU_HSVEB     STANDARD;      PRT;  3421 AA.
AC   P28955;
DT   01-DEC-1992 (Rel. 24, Created)
DT   01-DEC-1992 (Rel. 24, Last sequence update)
DT   01-APR-1993 (Rel. 25, Last annotation update)
DE   Large tegument protein.
GN   24.
OS   Equine herpesvirus type 1 (strain Ab4p) (EHV-1).
OC   Viruses; dsDNA viruses, no RNA stage; Herpesviridae;
OC   Alphaherpesvirinae; Varicellovirus.
OX   NCBI_TaxID=31520;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92295566; PubMed=1318606;
RA   Telford E.A.R., Watson M.S., McBride K., Davison A.J.;
RT   "The DNA sequence of equine herpesvirus-1.";
RL   Virology 189:304-316(1992).
CC   -!- FUNCTION: TEGUMENT PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY THAT GROUPS TOGETHER HSV-1 UL36,
CC       EHV-1 24, EBV BPLF1, HVS-1 64, VZV 22, AND HCMV UL48.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M86664; AAB02459.1; -.
DR   PIR; G36797; WZBEB6.
DR   HSSP; P04002; 1WFA.
SQ   SEQUENCE   3421 AA;  367078 MW;  5075EFE4739BB7AC CRC64;


alignment_scores:
           Quality:  117.00              Length:    591
             Ratio:  0.468                Gaps:     29
 Percent Similarity:  42.301   Percent Identity: 21.827

alignment_block:
 US-09-303-518D-465 x TEGU_HSVEB   ..

 Align seg 1/1  to: TEGU_HSVEB  from: 1  to: 3421

      98 TCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCAC 147
         ||||||         |||:::      :::   |||      ::::::||
    2585 SerGlyProProProThrLeuProProAlaProProLeuProGlnSerTh 2601

     148 CTATTCGGCAGCAGGGGGG.........AACTTGCCGAGCGCAGCGGTCA 188
         |   :::|||||||:::|||         :::||||||   |||      :
    2601 rSerLysAlaAlaSerGlyProProProThrLeuProProAlaProProL 2618

     189 TATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCC 238
         ::   :::   :::|||   ||||||   |||   ||||||   |||:::
    2618 euProGln...SerThrSerLysAlaAlaSerGlyAlaThrGlnSerAsp 2633

     239 AGCAGGCGGCCATTAAAG...GAAATATCGGCTACATTGTCC........ 277
         |||   :::   |||      ::::::      |||   |||
    2634 SerGlyLysThrLeuThrLeuAspValProLysThrGlnSerLysAspLy 2650

     278 .GCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGC 326
          :::   |||:::      |||   :::      ||||||||||||   |
    2650 sValValProValProProThrAspLys......ProSerThrThrThrP 2665

     327 CTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCC 376
         ||
    2665 ro.............................................. 2665

     377 TTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTAT 426
               |||::::::      :::      :::   ||||||||||||
    2666 ......AlaAlaLeuLysGlnSerAspAlaSerLysProProThrAlaAl 2680

     427 GACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATAT 476
            |||:::                                 :::   :::
    2680 aIleGlnHisGln.........................GlnLysLeuG 2688

     477 ATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGA 526
           |||   :::|||           ||||||   |||::::::
    2688 lyThrProValThr.............ProLysAspSerGlyAspLys... 2699

     527 CCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACC 576
         ||||||            |||||||::   :::   |||::::::   ||
    2700 ProThr.............AspAsnAlaSerAlaProValGlyValSerPr 2712

     577 GGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCG 626
         ||||:::            :::   :::      ::::::   ||||||
    2712 oValThr.........ProAspGlyThrProGlyAlaLysProProProL 2726
```

```
 627 ATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACG 676
2726 ysAspAlaProProValAspAspThrLysGlnProValArgLysSerLeu 2742

 677 GCACTGCAGATA.................................... 688
2743 ProSerGlnValArgGlyGlyArgProTyrIleArgProSerLeuGlyPr 2759

 688 ............................................... 688
2759 oPheLysPheThrGlyProProGlyTyrThrIleProValHisGlyLeuP 2776

 689 ..........TCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCG 727
2776 roProSerAspSerAsnValThrGlnSerThrLysGluProProLysPro 2792

 728 GCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTT 777
2793 AlaValGluThrProAlaAlaAlaProAlaLysSer.............. 2804

 778 ATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAA 827
2805 .AlaAlaAlaProAlaAlaAlaProAlaLysSer................ 2815

 828 CGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCC 877
2816 ................................AlaAlaAlaProAla 2820

 878 GCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTC 927
2821 AlaAlaProAlaLysSerAlaAlaAlaProAlaAlaAlaProAlaLysSe 2837

 928 AGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCG 977
2837 rAlaAlaAla..ProAlaAlaAlaProAlaLysSerAlaAlaAlaPro.. 2852

 978 GGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGC 1027
2853 .......................AlaAlaAlaProAlaLysAspGlnT 2861

1028 AGATGGGCGAGATCGCATTGCCGAAA...GGGAAATCCGCCGTCAGCGAC 1074
2861 hrLysSerAlaAlaGluValProLysProAlaLysAspGlnAlaLysAsp 2877

1075 AATTTTGCCGATGCGGCA......TACGCCAAATACCCGTCCCCTTACCA 1118
2878 GlnAlaLysAspGlnAlaLysAspGlnAlaLysAspGlnAla........ 2891

1119 TTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACA 1168
2892 .LysAspGlnAlaLysSerThrThrGlyGlnLysLeuAlaLysAsp.... 2906

1169 TCACCTCCTCAACCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCA 1218
2907 ................ProLysSerAspGlyLeuThrAspAspValAla 2917

1219 AACAAACGCCACCCG...AAGACCAAAGTGCCGTTTGACGGTAAAGGGTT 1265
2918 LeuGluIleValProGluLysThrProLeuProAspAspSerPro..... 2932

1266 TCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACCGCTG 1315
2933 .....................................IleGlyAlaV 2936

1316 TACCACAAGTGAATCCTATA......GATGAACCCGTCTTTAATCCTAAA 1359
2936 alProGluAsnThrProLeuProAspAspSerProIleGlySerProAsp 2952

1360 GGTTCTGTCGGA.........................TCGGCTCATTC 1382
2953 LeuSerAlaSerLysAsnSerHisThrThrAspAlaValSerSerAspAr 2969

1383 TTGGTCTATAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTA 1432
2969 gPheSerValAlaCysLysValProLeuProAspSerProGluAspAspP 2986

1433 GAATCAGATATATCCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCA 1482
2986 heTyr..................SerTyrAlaValAspValProLeuPro 2996

1483 AAAGGA............CCTAATAATGGATATTTGGATAAA........ 1512
2997 AspSerProThrAspAspProSerSerGlyArgSerAspAlaArgAlaPr 3013

1513 .............TTTGGTAATGAATGGACTAAAGGTCCATCAAGAACTA 1549
3013 oThrValGlyGlyValAlaSerIleHisArgLysSerAspSerArgAsnA 3030

1550 AAGGTCAAGAATTTGAATGGGATGTTCAATTGTCTAAAACA......GGA 1593
3030 snArgGlnSerAspAlaTrpArgArgAlaPheAlaAspThrLeuHisGly 3046

1594 AGAGAGCAACTTGGATGGGCTAGTAGGGATGGTAAGCATTTAAATATATC 1643
3047 ArgProArgAsnArgSerAlaThrLysProCysLys............Se 3059

1644 AATTGATGGAAAGATTACACAC 1665
3059 rAlaProTyrLysValProHis 3066
```

seq_name: SwissProt_40:MUC1_XENLA

seq_documentation_block:

```
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   Integumentary mucin C.1 (FIM-C.1) (Fragment).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P01359; 2PSP.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
```

```
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;

alignment_scores:
            Quality:  116.50            Length:    250
              Ratio:  0.971               Gaps:     12
 Percent Similarity:  48.000   Percent Identity: 25.600

alignment_block:

US-09-303-518D-465 x MUC1_XENLA  ..

Align seg 1/1  to: MUC1_XENLA  from: 1  to: 662

    110 TCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGC 159
        :::|||:::::            ||||||:::  |||  :::||||
      3 ThrThrAlaAlaAlaValAlaAlaThrGlyLysAspThrThrAlaAlaAl 19

    160 AGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACA 209
        |  |||:::     :::     :::  :::     |||:::  :
     19 aGluGlySerAlaAlaAlaGluLysThrAlaAlaAlaGlyGluVal...S 35

    210 AAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGGAA 259
        ::|||       |||:::  :::::::::  :::
     35 erAlaProProThrAlaAlaValAlaAlaThrGlyGluAspAlaThrThr 51

    260 ATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATTCC 309
           :::||||||  :::|||        :::::::::  :::  ||
     52 AlaAlaAlaThrAlaAlaAlaGluThrThrAlaAlaAlaGlyGluAlaPr 68

    310 CCCTTCGACAACCATGCCTCACATTCCGATTCTGATGAAGCCGGTAGTCC 359
        |  :::||||||  |||
     68 oThrThrThrThrAlaPro.............................. 74

    360 CGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAACACC 409
                     |||  ||||||:::             :::
     75 .............AlaThrThrAlaAla.............GlyLys 81

    410 ATCCCGCCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCC 459
           |||  ||||||              ||||||:::  |||
     82 AlaProThrThrAla.............AlaAlaThrAlaProThrTh 93

    460 AAAGGCGCGAGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAA 509
           |||  |||      |||::::::|||  ||||||  |||
     93 rAlaAlaAlaGlyAlaProThrThrAlaThrGlyLysAla...ProAlaT 109

    510 TATCCGCCTCAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCG 559
          :::|||:::      ||||||||||||||
    109 hrAlaAlaAlaProValProThrThrAlaAla................. 119

    560 ACCGTTTCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGA 609
              |||      |||:::::::::                 :::|||
    120 ......SerLysAlaProThrThrAlaAlaAlaAlaThrHisSerThrAl 134

    610 TTCAAACGCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGC 659
         :::  |||  |||  ||||||  ||| :::      ::::::
    134 aAlaAlaAlaAlaProThrThrAlaAlaSer.AlaAlaLysSerLysGlu 150

    660 CGC............CGAAGCTTTCAACGGCACTGCAGATATCGT...CA 694
        |||             |||  ::::::||||||:::  :::  ::
    151 ArgSerThrSerSerSerSerGluGluGluHisCysHisValLysProSe 167

    695 AAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTG 744
        :|||::::::      |||  ::::::      :::::::::|||:::
    167 rLysArgGluMetCysGlySerLysGlyIleThrLysLysGlnCysLysL 184

    745 CAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGG... 791
                          |||:::||||||:::         |||
    184 ys.................LysAsnCysCysPheAspProLysGlyHis 194

    792 .........TCTGCTTTCCACCGAAAACAAGATGGC...GCGCATCAA 827
                  :::||||||||||||      |||  :::|||:::
    195 GlyGlyIleHisCysPheHisArgLysProLysGlyHisSerHisGlu 210

seq_name: SwissProt_40:SUZ2_DROME

seq_documentation_block:
ID   SUZ2_DROME     STANDARD;      PRT;  1365 AA.
AC   P25172;
DT   01-MAY-1992 (Rel. 22, Created)
DT   01-MAY-1992 (Rel. 22, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Suppressor 2 of zeste protein (Protein posterior sex combs).
```

```
GN   SU(Z)2.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=91279476; PubMed=2057369;
RA   Brunk B.P., Adler P.N.;
RT   "The sequence of the Drosophila regulatory gene Suppressor two of
RT   zeste.";
RL   Nucleic Acids Res. 19:3149-3149(1991).
CC   -!- FUNCTION: REGULATES EXPRESSION OF THE HOMEOTIC SELECTOR GENES BY
CC       INFLUENCING HIGHER-ORDER CHROMATIN STRUCTURE THROUGH INTERACTION
CC       WITH OTHER PROTEINS.
CC   -!- SUBCELLULAR LOCATION: Nuclear (Probable).
CC   -!- SIMILARITY: CONTAINS 1 RING-TYPE ZINC FINGER.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X56798; CAA40134.1; -.
DR   EMBL; X56799; CAA40135.1; -.
DR   PIR; S16845; S16845.
DR   PIR; S14871; S14871.
DR   FlyBase; FBgn0008654; Su(z)2.
DR   InterPro; IPR001841; Znf_ring.
DR   Pfam; PF00097; zf-C3HC4; 1.
DR   SMART; SM00184; RING; 1.
DR   PROSITE; PS00518; ZF_RING_1; 1.
DR   PROSITE; PS50089; ZF_RING_2; 1.
KW   Zinc-finger; Developmental protein; DNA-binding; Nuclear protein.
FT   ZN_FING      35     74       RING-TYPE.
FT   DOMAIN      623    628       POLY-GLN.
FT   DOMAIN     1077   1096       POLY-ASN.
FT   DOMAIN     1241   1251       POLY-SER.
FT   CONFLICT    603    603       MISSING (IN REF. 1; CAA40134).
FT   CONFLICT    785    785       K -> N (IN REF. 1; CAA40134).
FT   CONFLICT    831    831       A -> R (IN REF. 1; CAA40134).
FT   CONFLICT    965    965       MISSING (IN REF. 1; CAA40134).
FT   CONFLICT   1065   1065       D -> E (IN REF. 1; CAA40134).
FT   CONFLICT   1076   1076       MISSING (IN REF. 1; CAA40134).
FT   CONFLICT   1287   1287       A -> P (IN REF. 1; CAA40134).
SQ   SEQUENCE   1365 AA;  146058 MW;  7B4BA0F35B0FA683 CRC64;


alignment_scores:
           Quality:  115.50            Length:    533
             Ratio:  0.509              Gaps:     22
 Percent Similarity:  42.589   Percent Identity: 20.075

alignment_block:
 US-09-303-518D-465 x SUZ2_DROME   ..

 Align seg 1/1  to: SUZ2_DROME  from: 1  to: 1365

      58 ATGCATGCACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGT 107
         :::||||||:::   |||:::   ::::::   ::::::      :::::
     518 LeuHisAlaGluIleSerSerGlnThrAspGlyLysMetLysValLysIl 534

     108 TCTCGACCGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCA 157
         :      :::   :::   :::   |||   |||    :::::|||
     534 eThrAlaLysProAsnHisLysLeuAspPheLys.ArgSerHisSerLeu 550

     158 GCAGGGGGGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATA 207
         |||:::|||:::|||                    |||   :::

     551 AlaSerGlyGluLeu.....................AspLeuGlnLysLe 560

     208 CAAAGCCATCAGTTGGGCAACCTGTTCATCCAGCAGGCGGCCATTAAAGG 257
          |||      |||   :::|||   ::::::   :::|||   |||   |
     560 uLysLeuAspSerThrSerThrSerGluAlaLeuAsnArgThrLeuGlyG 577

     258 AAATATCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCCATT 307
         ||   :::   ::::::::::::::   :::   |||:::   :::
     577 luGluAlaArgSerIleAsnSerLeuValValGlyGlyAlaProThrPro 593

     308 CCCCCTTCGACAACC................................. 322
         ||||||:::   |||
     594 ProProThrProThrAlaGluProGluGlnGlnGlnGlnGlnGlnGlnGl 610

     322 ............................................... 322

     610 nGlnGlnGlnProGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnGlnG 627

     323 .............ATGCCTCACATTCCGATTCTGATGAAGCCGGTA... 355
                     :::|||   |||      |||      |||:::
     627 lnGlnPheValValLeuProLysIleLysAspLeuThrLeuProThrSer 643

     356 GTCCCGTTGACGGATTCAGCCTTTACCGCATCCATTGGGACGGATACGAA 405
            ||||||      |||   |||   |||          :::   |||
     644 ProProLeuProProSerLeuPheLysAlaTyrThrProSerThrThrPr 660

     406 CACCATCCCGCCGACGGCTATGACGGGC.........CACAGGGCGGCGG 446
          |||   |||   |||:::      |||         ::::::
     660 oThrAlaProHisThrValAlaGlyGlyLysProLysGlnGlnGlnGlnG 677

     447 CTATCCCGCTCCCAAAGGCGCGAGGGATATAT......ACAGCTACGACA 490
           :::|||                  :::             |||   |||
     677 lnMetProGlnGlnProGlnAlaValLeuGlnGlnSerMetAlaLysThr 693

     491 TAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAACCGCAGC 540
               |||   |||   :::|||::::::         ::::::
     694 AsnProAlaLysProProLeuSerSerAsn.........AsnAsnArgLy 707

     541 ACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTATGCTGAC 590
          |||:::::::|||      :::|||        |||
     707 sProAsnSerGlyHisPheAlaValProGlnAlaPro.............. 719

     591 GCAAGGAGTAGGCGACGGATTCAAAC.......................... 616
                       |||      |||
     720 .............ThrHisArgAsnMetTyrHisMetGlnArgTyrGln 731

     617 GCGCCACCCGATACAGCCCCGAGCTGGACAGATCGGGCAATGCCGCCGAA 666
         :::   |||::::::      |||   :::::::::          |||||
     732 SerThrProSerSerIleAlaSerAlaAlaAsnLysMet......ProLy 746

     667 GCTTTCAACGGCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGG 716
         |   |||   :::|||:::   |||      :::::::::||||||   :
     746 sArgSerMetSerLeuAspGluSerHisProAlaLysGlnAlaArgLeuS 763

     717 AGAAATTGTCGGCGCAGGCG.......................ATGCCG 742
         :::::   :::|||:::|||                        |||
     763 erGlnAlaGlnAlaMetAlaSerSerTyrAlaAlaLysLeuHisMetGln 779

     743 TGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGT 792
                 ||||||:::|||:::                       ::
     780 ThrSerAsnGlnAlaLysSerGlnAla.....................Al 789

     793 CTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATAT 842
         ::::|||   |||:::   :::   |||:::        :::   ::::::
     789 aAlaPheLeuProAsnProGlnMetArgSerTyrGlyLeuProAspLeu. 805

     843 GGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGCAGTCC 892
               ::::::   |||:::|||:::         |||   :::::|||
     806 ..GlySerLysProThrLeuProMetLeuCysProAlaSerSerSerSer 821

     893 AAA.........ACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAAT 933
         :::          ||||||   |||:::   :::      ||||||
     822 GlnValThrIleThrProArgProArgAlaThr......ProSerIleTy 836
```

```
     934 ATCTTTA..................................... 940
         |||:::
     836 rSerPheSerGluProAsnIleHisValProAlaLeuGluIleValArgL 853

     941 ..........CGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGG 979
                  :::||||||:::   :::||||||||||   :::
     853 euProValAsnLysGlnSerAlaGlyGlyLysGlyLeuThrMetProPro 869

     980 GAAAATACGGCTTGGGCGGCATCACGGCACATCCTG.............. 1015
         :::          :::::|||   :::::::::
     870 LeuSerProProAlaThrSerSerAlaArgLeuMetGlyProProAlaAl 886

    1016 ................TCAAGCGGTCGCAGATGGGCGAGATCGCATTGCC 1049
                        :::   |||:::      |||:::      |||:
     886 aLeuProLysHisAlaGlyHisGlyHisGlyAlaAlaLysArgSerCysG 903

    1050 GAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATGCGGCATACGCCA 1099
         ::      :::   |||            ||||||:::
     903 lnMetProThrMetProMetProLeuProLeuProLeuProMetProMet 919

    1100 AATACCCGTCCCCTTACCATT......CCCGAAATATCCGTTCAA..... 1138
         :::|||   |||      :::      |||   :::||||||
     920 ThrThrIleProAlaIleValLysSerProProLeuSerValAlaLeuSe 936

    1139 .....................ACTTGGAGCAGCGTTACGGCAAAGAAAA 1166
                              :::   |||:::   :::|||   :::|
     936 rGlyGlnArgAsnAsnLysGlyAsnSerSerAsnSerAsnAlaTyrArgT 953

    1167 CATCACCTCCTCAA................................. 1180
         ||||||||||
     953 hrSerProProAlaLeuIleAsnLeuArgAsnThrAlaAlaProGlnHis 969

    1181 ......CCGTGCCGCCGTCAAACGGAAAGAATGTGAAACTGGCAAACAAA 1224
               |||   :::   :::   :::                :::::||
     970 SerPheProSerLysSerSerProLys............ValGluAlaAs 982

    1225 CGCCACCCGAAGACCAAAG.........TGCCGTTTGACGGTAAAGGGTT 1265
         |:::   :::   |||           |||:::      :::::|||
     982 nSerLysLysSerProProAlaAlaGlyCysGlnGlyLysThrAsnGlyT 999

    1266 TCCGAATTTTGAAAAAGACGTAAAATACGATACGAGAATTAATACCG 1312
                 |||   |||:::   :::::::::||||||:::   |||
     999 hrAlaAlaLeuAspLysSerLysThrSerLeuArgGluPheArgPro 1014

seq_name: SwissProt_40:OMPA_RICCN

seq_documentation_block:
ID   OMPA_RICCN     STANDARD;      PRT;  2021 AA.
AC   Q52657; P95591; P95592; P95593; P95594; Q52667; Q52668; Q52669;
AC   Q52670; Q52674;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Outer membrane protein A precursor (190 kDa antigen) (Cell surface
DE   antigen) (rOmpA) (rOmp A).
GN   OMPA OR RC1273.
OS   Rickettsia conorii.
OC   Bacteria; Proteobacteria; alpha subdivision; Rickettsiales;
OC   Rickettsiaceae; Rickettsieae; Rickettsia.
OX   NCBI_TaxID=781;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=Malish 7;
RX   MEDLINE=94171067; PubMed=8125327;
RA   Crocquet-Valdes P.A., Weiss K., Walker D.H.;
RT   "Sequence analysis of the 190-kDa antigen-encoding gene of Rickettsia
RT   conorii (Malish 7 strain).";
RL   Gene 140:115-119(1994).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=Malish 7;
RX   MEDLINE=21442074; PubMed=11557893;
RA   Ogata H., Audic S., Renesto-Audiffren P., Fournier P.-E., Barbe V.,
RA   Samson D., Roux V., Cossart P., Weissenbach J., Claverie J.-M.,
RA   Raoult D.;
RT   "Mechanisms of evolution in Rickettsia conorii and R. prowazekii.";
RL   Science 293:2093-2098(2001).
RN   [3]
RP   SEQUENCE OF 8-204 FROM N.A.
RC   STRAIN=Indian tick typhus, M1, Malish 7, and Moroccan;
RX   MEDLINE=97015921; PubMed=8862558;
RA   Roux V., Fournier P.E., Raoult D.;
RT   "Differentiation of spotted fever group rickettsiae by sequencing and
RT   analysis of restriction fragment length polymorphism of PCR-amplified
RT   DNA of the gene encoding the protein rOmpA.";
RL   J. Clin. Microbiol. 34:2058-2065(1996).
RN   [4]
RP   SEQUENCE OF 953-2012 FROM N.A.
RC   STRAIN=Indian tick typhus, M1, Malish 7, and Moroccan;
RA   Raoult D., Fournier P.E., Roux V.;
RT   "Phylogenetic analysis of spotted fever group rickettsiae by study
RT   of the outer surface protein rOmpA.";
RL   Submitted (DEC-1996) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: ELICITS PROTECTIVE IMMUNITY (BY SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: CELL WALL. THIS BACTERIUM IS COVERED BY A
CC       S-LAYER WITH HEXAGONAL SYMMETRY.
CC   -!- PTM: GLYCOSYLATED (BY SIMILARITY).
CC   -!- SIMILARITY: BELONGS TO THE RICKETTSIAE OMPA/OMPB FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U01028; AAA17405.1; -.
DR   EMBL; AE008674; AAL03811.1; -.
DR   EMBL; U43794; AAB49549.1; -.
DR   EMBL; U43798; AAB49550.1; -.
DR   EMBL; U43806; AAB49551.1; -.
DR   EMBL; U45244; AAB49566.1; -.
DR   EMBL; U46918; AAA86663.1; -.
DR   EMBL; U83440; AAC35176.1; -.
DR   EMBL; U83443; AAC35179.1; -.
DR   EMBL; U83448; AAC35184.1; -.
DR   EMBL; U83453; AAC35189.1; -.
DR   InterPro; IPR003858; rOmpA_rOmpB.
DR   Pfam; PF02708; rOmpA_rOmpB; 1.
KW   Antigen; Repeat; Signal; Cell wall; S-layer; Glycoprotein;
KW   Complete proteome.
FT   SIGNAL       1     38       POTENTIAL.
FT   CHAIN       39   2021       OUTER MEMBRANE PROTEIN A.
FT   DOMAIN     238    946       THR-RICH.
FT   DOMAIN    1424   1528       THR-RICH.
FT   VARIANT     60     60       N -> NN (IN STRAIN INDIAN TICK TYPHUS).
FT   VARIANT     76     76       R -> H (IN STRAIN INDIAN TICK TYPHUS).
FT   VARIANT     86    137       MISSING (IN STRAIN M1).
FT   VARIANT    126    133       MISSING (IN STRAIN MOROCCAN).
FT   VARIANT    953    954       VT -> II (IN STRAIN INDIAN TICK TYPHUS).
FT   VARIANT   1245   1245       D -> A (IN STRAINS INDIAN TICK TYPHUS, M1
FT                               AND MOROCCAN).
FT   VARIANT   1308   1308       N -> H (IN STRAIN MOROCCAN).
FT   VARIANT   1877   1877       M -> I (IN STRAIN INDIAN TICK TYPHUS).
FT   CONFLICT    10     10       Q -> K (IN REF. 1).
FT   CONFLICT    92     92       I -> V (IN REF. 1).
FT   CONFLICT   126    126       V -> I (IN REF. 1).
FT   CONFLICT   137    137       T -> N (IN REF. 1).
FT   CONFLICT   157    157       G -> D (IN REF. 1).
FT   CONFLICT   368    369       IS -> VN (IN REF. 1).
FT   CONFLICT   374    388       KATLGGAIIKATTTK -> LLQVQGGVVKANTIN (IN
FT                               REF. 1).
FT   CONFLICT   640    640       N -> D (IN REF. 1).
FT   CONFLICT   669    669       V -> I (IN REF. 1).
```

```
FT   CONFLICT    793    793       N -> D (IN REF. 1).
FT   CONFLICT    803    804       VN -> IS (IN REF. 1).
FT   CONFLICT    809    823       LLRVQGGVVKSNTIN -> KATLGGAIIKATTTK (IN
FT                                REF. 1).
FT   CONFLICT    898    898       D -> Y (IN REF. 1).
FT   CONFLICT    908    908       P -> N (IN REF. 1).
FT   CONFLICT    985    985       N -> K (IN REF. 1).
FT   CONFLICT   1009   1009       L -> S (IN REF. 1).
FT   CONFLICT   1013   1013       Y -> S (IN REF. 1).
FT   CONFLICT   1182   1182       K -> Q (IN REF. 1).
FT   CONFLICT   1314   1314       N -> Y (IN REF. 4).
FT   CONFLICT   1451   1451       H -> N (IN REF. 1).
FT   CONFLICT   1624   1624       G -> D (IN REF. 1).
FT   CONFLICT   1628   1628       E -> G (IN REF. 1).
FT   CONFLICT   1872   1872       A -> V (IN REF. 1).
FT   CONFLICT   1875   1875       T -> P (IN REF. 1).
FT   CONFLICT   1878   1879       MS -> LP (IN REF. 1).
FT   CONFLICT   1936   1936       E -> A (IN REF. 1).
FT   CONFLICT   1965   1970       MTAPLP -> ITPPLS (IN REF. 1).
FT   CONFLICT   1997   1997       G -> R (IN REF. 1).
SQ   SEQUENCE   2021 AA;  203328 MW;  327FC42D7CB24668 CRC64;


alignment_scores:
           Quality:  118.00            Length:    384
             Ratio:  0.648               Gaps:     15
 Percent Similarity:  47.396   Percent Identity: 21.875

alignment_block:
 US-09-303-518D-465 x OMPA_RICCN   ..

 Align seg 1/1  to: OMPA_RICCN  from: 1  to: 2021

      86 ACGATTCTTTTATCCGGCAGGTTCTCGACCGTCAGCATTTCGAACCCG.. 133
         |||      |||:::   :::   |||:::|||:::::::::|||||  
     386 ThrThrLysLeuThrAspAsnAlaSerAlaValThrPheThrAsnProVa 402

     133 ..................................................... 133

     402 lValValThrGlyAlaIleAspAsnThrGlyAsnAlaAsnAsnGlyIleV 419

     134 ........ACGGGAAATACCACCTATTCGGCAGCAGGGGGGAACTTGCCG 175
                 ||||||::::::|||   ::::::     ||||||
     419 alThrPheThrGlyAspSerThrValThrGlyAsnIleGlyAsnThrAsn 435

     176 AGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGCCATCAGTTGGGC 225
         :::   |||:::||||||              :::::::::   ||
     436 AlaLeuAlaThrIleSerValGlyAlaGlyLysAlaThrLeuGlyGlyAl 452

     226 AACCTGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATTGT 275
         |      :::::::::         |||                |||   :
     452 aIleIleLysAlaThrThrThrLysLeu..............ThrAspA 464

     276 CCGCTTTTCCGATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATG 325
         ::|||     :::|||   |||:::   :::      :::   ::::::
     464 snAlaSerAlaValThrPheThrAsnProValValValThrGlyAlaIle 480

     326 CCTCACATTCCGATTCTGATGAAGCCGGTAGTCCCGTTGACG...GATTC 372
            :::               :::   :::|||   ::::||   |||||
     481 AspAsnThrGlyAsnAlaAsnAsnGlyIleValThrPheThrGlyAspSe 497

     373 AGCCTTTACCGCATCCATTGGGACGGATACGAACACCATCCCGCCGACGG 422
         |:::   |||::::::|||||   :::|||||:::::         :
     497 rThrValThrGlyAsnIleGly...AsnThrAsnAlaLeuAlaThrIleS 513

     423 CTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGG 472
         :::::   :::   :::|||:::                :::   :::
     513 erValGlyAlaGlyLysAlaThrLeu..............GlyGlyAla 524

     473 ATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAAC 522
         |||      |||||||||   |||         :::   ||||||   ||
     525 IleIleLysAlaThrThrThrLysLeuThrAspAsnAlaSerAlaValTh 541

     523 CTGACCG...........ACAACCGCAGCACCGGACAAC......... 550
         |                 :::|||:::|||   ||||||
     541 rPheThrAsnProValValValThrGlyAlaIleAspAsnThrGlyAsnA 558

     551 .......................GGCTTGTCGACCGTTTCCAC.... 570
                                |||   |||||||||:::
     558 laAsnAsnGlyIleValThrPheThrGlyAspSerThrValThrGlyAsn 574

     571 ......AATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGA..... 609
               ||||||::::::::   |||   :::||||||   |||
     575 IleGlyAsnThrAsnAlaLeuAlaThrIleSerValGlyAlaGlyLysAl 591

     609 ..................................................... 609

     591 aThrLeuGlyGlyAlaIleIleLysAlaThrThrThrLysLeuThrAspA 608

     610 ................TTCAAACGCGCCACCCGATACAGCCCCGAGCTG 642
                        |||   :::   :::      :::      :::
     608 snAlaSerAlaValThrPheThrAsnProValValValThrGlyAlaIle 624

     643 GACAGATCGGGCAATGCCGCCGAA.........GCTTTCAACGGCACTGC 683
         |||::::::|||||||||   :::          :::|||:::|||:::::
     625 AspAsnThrGlyAsnAlaAsnAsnGlyIleValThrPheThrGlyAsnSe 641

     684 AGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAG 733
         :   ::::::   ||||||                          |||
     641 rThrValThrGlyAsnIle.......................GlyAsnT 650

     734 GCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTTATGCAC 783
           :::|||:::      ::::::   |||::::::||||||:::::::::
     650 hrAsnAlaLeuAlaThrValAsnValGlyAlaGlyIleAlaThrLeuGlu 666

     784 GGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAACGATTT 833
         |||      :::   :::|||   :::|||:::::::::     :::::
     667 GlyAlaValIleLysAlaThrThrThrLysLeuThrAsnAlaAlaSerVa 683

     834 GGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATT 883
         :      :::::::::::::::               |||:::
     683 lLeuThrLeuThrAsnValAsn..............AlaValLeuThrG 695

     884 GGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAAT 933
           |||::::::::|||   :::::::::::::::::   :::|||   |||
     695 lyAlaIleAspAsnThrThrGlyValAspAsnVal...GlyValLeuAsn 710

     934 ATCTTTACGGCAGTCATCCCCGTCAAAGGG..................... 963
         :::      |||:::      |||   |||
     711 LeuAsnGlyAlaLeuSerGlnValThrGlyAsnIleGlyAsnThrAsnAl 727

     964 ..............ATTGGAGCTGTTCGGGGAAAATACGGCTTGGGCG 997
                        :::||||||      ||||||      |||||||
     727 aLeuAlaThrIleSerValGlyAla......GlyLysAlaThrLeuGlyG 742

     998 GC 999
         ||
     742 ly 742

seq_name: SwissProt_40:LG1_MAIZE

seq_documentation_block:
ID   LG1_MAIZE      STANDARD;      PRT;   399 AA.
AC   O04003;
DT   15-JUL-1999 (Rel. 38, Created)
DT   15-JUL-1999 (Rel. 38, Last sequence update)
DT   15-JUL-1999 (Rel. 38, Last annotation update)
DE   LIGULELESS1 protein.
GN   LG1.
OS   Zea mays (Maize).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACC clade;
OC   Panicoideae; Andropogoneae; Zea.
OX   NCBI_TaxID=4577;
RN   [1]
RP   SEQUENCE FROM N.A.
```

```
RC   STRAIN=CV. W22; TISSUE=Leaf;
RX   MEDLINE=97230364; PubMed=9119226;
RA   Moreno M.A., Harper L.C., Krueger R.W., Dellaporta S.L., Freeling M.;
RT   "Liguleless encodes a nuclear-localized protein required for
RT   induction of ligules and auricles during maize leaf organogenesis.";
RL   Genes Dev. 11:616-628(1997).
CC   -!- FUNCTION: INVOLVED IN THE FORMATION OF LIGULES AND AURICLES DURING
CC       LEAF ORGANOGENESIS.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- TISSUE SPECIFICITY: LEAF LIGULAR REGION, BLADE AND SHEATH.
CC   -!- DEVELOPMENTAL STAGE: DEVELOPING LEAVES. PRESENT AT HIGHEST
CC       LEVELS IN VERY YOUNG LIGULES AND AURICLES. PRESENT PRE- AND POST-
CC       LIGULE AND AURICLE DEVELOPMENT.
CC   -!- SIMILARITY: CONTAINS 1 SBP DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U89496; AAB51071.1; -.
DR   Mendel; 17084; ZEAma;1377;mn17084.
DR   InterPro; IPR004333; SBP.
DR   Pfam; PF03110; SBP; 1.
KW   Nuclear protein; Developmental protein.
FT   DOMAIN      243    259       NUCLEAR LOCALIZATION SIGNAL (POTENTIAL).
FT   DOMAIN      185    260       SBP.
SQ   SEQUENCE   399 AA;  43367 MW;  1E89F304232D7B8C CRC64;


alignment_scores:
           Quality:  114.50            Length:    374
             Ratio:  0.666              Gaps:     20
 Percent Similarity:  45.989   Percent Identity: 22.995

alignment_block:
 US-09-303-518D-465 x LG1_MAIZE   ..

 Align seg 1/1  to: LG1_MAIZE  from: 1  to: 399

     140 AATACCACCTATTCGGCAGCAGGGGGGAACTTGCCGAGCGCAGCGGTCAT 189
         :::||||||::       ||||||||::   |||       :::
      73 HisThrThrPhe......AlaGlyGlyHisSerProHisAspIleLeuHi 87

     190 ATCGGATTGGGAAACATACAAAGCCATCAGTTGGGCAACCTGTTCATCCA 239
                        :::|||   |||         :::::::::
      87 sPheTyrThrProProProSerAlaAlaSerHisTyrLeuAlaAlaAlaA 104

     240 GCAGGCGGCCATTAAAGGAA...ATATCGGCTACATTGTCCGCTTTTCCG 286
         ::   :::|||   ::::::   :::||||||      ::::::|||
     104 laGlyAsnProTyrSerHisLeuValSerAlaProGlyThrThrPheHis 120

     287 ATCACGGGCACGAAGTCCATTCCCCCTTCGACAACCATGCCTCACATTCC 336
           |||::::::::::      ||||||:::::::::          ||
     121 GlnThrSerSerSerTyrTyrProProAlaAlaAlaAlaGlnAlaAlaPr 137

     337 GATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCC.......... 376
         |          |||:::      |||::::::||||||
     137 oGluTyrTyrPheProThr......LeuValSerSerAlaGluGluAsnM 152

     377 ........TTTACCGCATCC...ATTGGGACGGATACGAACACCATCCCG 415
                 |||:::|||:::   :::|||   :::
     152 etAlaSerPheAlaAlaThrGlnLeuGlyLeuAsn.............. 163

     416 CCGACGGCTATGACGGGCCACAGGGCGGCGGCTATCCCGCTCCCAAAGGC 465
                       |||:::|||:::            :::|||    |
     164 ............LeuGlyTyrArgThrTyr.........ProProPro.A 173

     466 GCGAGGGATATATACAGCTACGACATAAAAGGCGTTGCCCAAAATATCCG 515
         ||   ||||||         :::|||      ||||||

     173 rgGlyGlyTyrThrTyrGlyHisHisProProArgCys............ 185

     516 CCTCAACCTGACCGACAACCGCAGCACCGGACAACGGCTTGTCGACCGTT 565
                                      ::::::   :::|||:::
     186 ..........................GlnAlaGluGlyCysLys..... 191

     566 TCCACAATACCGGTAGTATGCTGACGCAAGGAGTAGGCGACGGATTCAAA 615
                                 ||||||      |||   :::|||   :::
     192 ..................AlaAspLeuSerSerAlaLysArgTyrHisA 202

     616 CGCGCCAC..........................CCGATACAGCCCCGA 638
           ||||||                          |||:::
     202 rgArgHisLysValCysGluHisHisSerLysAlaProValValValThr 218

     639 GCTGGACAGATCGGGCAA......TGCCGCCGAAGCTTTCAACGGCAC.. 680
         ||||||   :::   |||      |||::::::      :::   |||
     219 AlaGlyGlyLeuHisGlnArgPheCysGlnGlnCysSerArgPheHisLe 235

     681 ..............................TGCAGATATCGTCAAAAAC 699
                                       ||||||   |||
     235 uLeuAspGluPheAspAspAlaLysLysSerCysArgLysArgLeuAlaA 252

     700 ATCATCGGCGCGGCAGGAGAAATTGTCGGCGCAGGCGATGCCGTGCAGGG 749
           |||:::|||   ||||||:::   :::         :::   ||||||
     252 spHisAsnArgArgArgArgLysSerLysProSerAspAlaAspAlaGly 268

     750 TATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTT 799
           |||:::|||
     269 AspLysLysArgAla................................... 273

     800 CCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCG... 846
          |||    ||||||   |||      :::|||   |||:::   :::
     274 .His.AlaAsnLysAlaAlaAlaAlaLysAspLysAlaGluSerSerSer 289

     847 .........CAACTCAAAGACTATGCCGCAGCAGCCATCCGCGATTGGGC 887
                   ::::::   |||      :::|||   |||         ||
     290 LysAsnMetAspIleGlyAspGlyLeuGlyAlaGlnIleLeuGlySerAl 306

     888 AGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTCAGCAATATCT 937
         |:::   :::   :::       |||   ::::::      :::   :::
     306 aLeuLeuSerLysGluGlnAspGlnThrMetAspLeuGlyGluValValL 323

     938 TTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGAAAATAC 987
               ||||||   |||   ||||||   ::::::::::::      |||
     323 ysGluAlaValAspPro...LysGlyLysAlaSerMetGlnGlnHisTyr 338

     988 GGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGCAGATGGGCGA 1037
         |||:::                  |||         |||:::   |||::
     339 GlyPhe.................ProPheHisSerSerSerAlaGlySe 349

    1038 GATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAATTTTGCCGATG 1087
         :   ::::::|||:::   :::::::::   ||||||:::   ::::::
     349 r...CysPheProGlnThrGlnAlaValSerSerAspThrThrSerAsnI 365

    1088 CGGCATACGCCAAATACCCGTCCCCT......TACCATTCCCGAAATATC 1131
           :::   ::::::   ||||||         :::|||::::::::::
     365 leGlyGlnValGlnGluProSerLeuGlyPheHisHisGlnHisHisGln 381

    1132 CGTTCAAACTTGGAGCAG 1149
         :::|||||||:::   |||
     382 HisSerAsnIleLeuGln 387

seq_name: SwissProt_40:MAP4_MOUSE

seq_documentation_block:
ID   MAP4_MOUSE     STANDARD;      PRT;  1125 AA.
AC   P27546;
DT   01-AUG-1992 (Rel. 23, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Microtubule-associated protein 4 (MAP 4).
GN   MAP4 OR MTAP4.
OS   Mus musculus (Mouse).
```

```
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92042100; PubMed=1718985;
RA   West R.R., Tenbarge K.M., Olmsted J.B.;
RT   "A model for microtubule-associated protein 4 structure. Domains
RT   defined by comparisons of human, mouse, and bovine sequences.";
RL   J. Biol. Chem. 266:21886-21896(1991).
CC   -!- FUNCTION: NON-NEURONAL MICROTUBULE-ASSOCIATED PROTEIN. PROMOTES
CC       MICROTUBULE ASSEMBLY.
CC   -!- TISSUE SPECIFICITY: TESTIS, STRIATED AND CARDIAC MUSCLE.
CC   -!- PTM: PHOSPHORYLATION OF THE PRO-RICH REGION IN THE C-TERMINUS
CC       NEGATIVELY REGULATES MAP-4 ACTIVITY TO PROMOTE MICROTUBULE
CC       ASSEMBLY.
CC   -!- SIMILARITY: CONTAINS 3 TAU/MAP REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M72414; AAA16372.1; -.
DR   PIR; B41206; B41206.
DR   MGD; MGI:97178; Mtap4.
DR   InterPro; IPR001084; Tubulin-bind.
DR   Pfam; PF00418; tubulin-binding; 4.
DR   PROSITE; PS00229; TAU_MAP; 3.
KW   Microtubules; Repeat; Phosphorylation.
FT   REPEAT      896    926       TAU/MAP MOTIF.
FT   REPEAT      965    995       TAU/MAP MOTIF.
FT   REPEAT      996   1026       TAU/MAP MOTIF.
SQ   SEQUENCE   1125 AA;  117675 MW;  73047432A329AA1D CRC64;


alignment_scores:
            Quality:  114.50            Length:    459
              Ratio:  0.581              Gaps:     21
 Percent Similarity:  42.919   Percent Identity: 21.786

alignment_block:
 US-09-303-518D-465 x MAP4_MOUSE   ..

 Align seg 1/1  to: MAP4_MOUSE  from: 1  to: 1125

     230 TGTTCATCCAGCAGGCGGCCATTAAAGGAAATATCGGCTACATTGTCCGC 279
         |||::::::|||   :::|||::::|||      :::      :::|||::
     595 CysThrAlaSerProGluProValLys...AlaAlaGluGlnMetSerTh 610

     280 TTTTCCGATCACGGGC.....................ACGAAGTCCA 305
         ::::||||||   :::                      |||:::
     610 rLeuProIleAspAlaProSerProLeuGluAsnLeuGluGlnLysGluT 627

     306 TTCCCCCTTCGACAACCATGCCTCACATTCCGATTCTG............ 343
           |||   |||      :::|||      |||
     627 hrProGlySer......GlnProSerGluProCysSerGlyValSerArg 641

     344 .........ATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCGC 384
                    |||   :::|||   :::|||   :::   :::|||     .
     642 GlnGluGluAlaLysAlaAlaValGlyValThrGlyAsnAspIleThr.. 657

     385 ATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGCC 434
                 |||      |||      ||||||
     658 .........ThrProProAsnLysGluProPro.............. 665

     435 ACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGCT 484
                       |||   |||:::   ::::::      |||
     666 ................ProSerProGluLysLysAlaLysProLeuAla 676

     485 ACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAAC 534
         ||||||      |||      :::   |||   :::            ||
     677 ThrThrGlnProAlaLysThrSerThrSerLysAlaLysThrGlnProTh 693

     535 CGCAGCACCGGACAACGGCTTGTCGACCGTTTCC.............. 568
         |:::   |||   :::         |||:::|||
     693 rSerLeuProLysGlnProAlaProThrThrSerGlyGlyLeuAsnLysL 710

     569 ......................ACAATACCGGTAGTATGCTGACGCAAG 595
                                 ::::::|||::::::      ::::::
     710 ysProMetSerLeuAlaSerGlySerValProAlaAlaProHisLysArg 726

     596 GAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGCCCCGAGCTGGAC 645
               |||:::   :::::::|||   |||:::|||   |||:::
     727 ProAlaAlaAlaThrAlaThrAlaArgProSerThrLeuProAla..... 741

     646 AGATCGGGCAATGCCGCCGAAGCTTTCAACGGCACTGCAGATATCG.... 691
           |||   :::   ||||||      |||      :::::::::
     742 ....ArgAspValLysProLysProIleThrGluAlaLysValAlaGluL 757

     692 .............TCAAAAACATCATCGGCGCGGCAGGAGAAATTGTCG 727
                       ||||||   |||||||||         |||   :::
     757 ysArgThrSerProSerLysProSerSerAlaProAlaLeuLysProGly 773

     728 GCGCAGGCGATGCCGTGCAGGGTATAAGCGAAGGCTCAAACATTGCTGTT 777
           ::::::   |||            |||   :::   :::   |||::
     774 ProLysThrThrProThrValSerLysAlaThrSerProSerThrLeuVa 790

     778 ATGCACGGCTTGGGTCTGCTTTCCACCGAAAACAAGATGGCGCGCATCAA 827
         :   |||:::                |||::::::|||      |||:::
     790 lSerThrGly..............ProSerSerArgSerProAlaThr. 801

     828 CGATTTGGCAGATATGGCGCAACTCAAAGACTATGCCGCAGCAGCCATCC 877
                                    |||:::|||::::::|||:::
     802 ............................ThrLeuProLysArgProThr 808

     878 GCGATTGGGCAGTCCAAAACCCCAATGCCGCACAAGGCATAGAAGCCGTC 927
         :::|||      :::      |||         |||
     809 SerIleLysThrGluGlyLysProAlaAspValLys.............. 820

     928 AGCAATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCG 977
                   |||::::::::::   |||   :::   :::|||
     821 .............ArgMetThrAlaLysSerAlaSerAlaAspLeu.... 831

     978 GGGAAAATACGGCTTGGGCGGCATCACGGCACATCCTGTCAAGCGGTCGC 1027
                              ||||||      :::|||:::
     832 ......................SerArgSerLysThrThrSerAlaSer 840

    1028 AGATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCGTCAGCGACAAT 1077
               |||:::           |||      |||   |||   ::
     841 SerValLysArgAsnThrThrProThrGlyAlaAlaProProAlaGlyMe 857

    1078 TTTGCCGATGCGGCATACGCCAAATACCCGTCCC.............. 1111
         :         |||      |||   ::   |||
     857 tThrSerThrArgValLysProMetSerAlaProSerArgSerSerGlyA 874

    1112 ..CTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCA 1159
           |||::::::   :::      ::::::|||      |||   :::|||
     874 laLeuSerValAspLysLysProThrSerThrLysProSerSerSerAla 890

    1160 AAGAAAACATCACC....TCCTCAACCGTGCCGCCGTCAAACGGAAAGAA 1205
           ::::::|||      ::::::||||||          :::   |||::
     891 ProArgValSerArgLeuAlaThrThrValSerAlaProAspLeuLysSe 907

    1206 TGTGAAA..............CTGGCAAACAAACGCCACCCGAAGACCA 1240
         :|||:::                     |||   :::|||
     907 rValArgSerLysValGlySerThrGluAsnIleLysHisGln....... 921

    1241 AAGTGCCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAA 1290
                   |||   |||      :::   ||||||      |||
     922 ...........ProGlyGlyGlyArgAlaLysValGluLys......Lys 932

    1291 TACGATACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGA 1340
```

```
          ::::::      :::::::::   |||:::   |||   :::   ::
     933 ThrGluAlaAlaThrThrAlaGlyLysProGluProAsnAlaValThrLy 949

    1341 ACCCGTCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTA 1390
         :   :::           ||||||::::::||||||::::::
     949 sAlaAla.............GlySerIleAlaSerAlaGlnLys....... 959

    1391 TAACTGCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGA 1440
                                    |||      |||:::::::::
     960 ............................ProProAlaGlyLysValGln 966

    1441 TATATCCCACCTAAA...AATTACTCT 1464
          :::      |||   :::||||||
     967 IleValSerLysLysValSerTyrSer 975

seq_name: SwissProt_40:NCR2_HUMAN

seq_documentation_block:
ID   NCR2_HUMAN     STANDARD;      PRT;  2517 AA.
AC   Q9Y618; Q9Y5U0; Q13354; O00613; O15416;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Nuclear receptor co-repressor 2 (N-CoR2) (Silencing mediator of
DE   retinoic acid and thyroid hormone receptor) (SMRT) (SMRTe) (Thyroid-,
DE   retinoic-acid-receptor-associated co-repressor) (T3 receptor-
DE   associating factor) (TRAC) (CTG26).
GN   NCOR2.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A. (ISOFORM SMRT).
RC   TISSUE=Pituitary;
RX   MEDLINE=99178941; PubMed=10077563;
RA   Ordentlich P., Downes M., Xie W., Genin A., Spinner N.B., Evans R.M.;
RT   "Unique forms of human and mouse nuclear receptor corepressor SMRT.";
RL   Proc. Natl. Acad. Sci. U.S.A. 96:2639-2644(1999).
RN   [2]
RP   SEQUENCE FROM N.A. (ISOFORM SMRT).
RC   TISSUE=Cervical adenocarcinoma;
RX   MEDLINE=99199215; PubMed=10097068;
RA   Park E.J., Schroen D.J., Yang M., Li H., Li L., Chen J.D.;
RT   "SMRTe, a silencing mediator for retinoid and thyroid hormone
RT   receptors-extended isoform that is more related to the nuclear
RT   receptor corepressor.";
RL   Proc. Natl. Acad. Sci. U.S.A. 96:3519-3524(1999).
RN   [3]
RP   SEQUENCE OF 1023-2517 FROM N.A.
RC   TISSUE=Cervical adenocarcinoma;
RX   MEDLINE=96008552; PubMed=7566127;
RA   Chen J.D., Evans R.M.;
RT   "A transcriptional co-repressor that interacts with nuclear hormone
RT   receptors.";
RL   Nature 377:454-457(1995).
RN   [4]
RP   SEQUENCE FROM N.A. (ISOFORM TRAC-1).
RC   TISSUE=Fetal liver;
RX   MEDLINE=96408715; PubMed=8813722;
RA   Sande S., Privalsky M.L.;
RT   "Identification of TRACs (T3 receptor-associating cofactors), a family
RT   of cofactors that associate with, and modulate the activity of,
RT   nuclear hormone receptors.";
RL   Mol. Endocrinol. 10:813-825(1996).
RN   [5]
RP   SEQUENCE OF 428-613 FROM N.A.
RC   TISSUE=Brain cortex;
RX   MEDLINE=97369492; PubMed=9225980;
RA   Margolis R.L., Abraham M.R., Gatchell S.B., Li S.H., Kidwai A.S.,
RA   Breschel T.S., Stine O.C., Callahan C., McInnis M.G., Ross C.A.;
RT   "cDNAs with long CAG trinucleotide repeats from human brain.";
RL   Hum. Genet. 100:114-122(1997).
CC   -!- FUNCTION: MEDIATES THE TRANSCRIPTIONAL REPRESSION ACTIVITY OF SOME
CC       NUCLEAR RECEPTORS BY PROMOTING CHROMATIN CONDENSATION, THUS
CC       PREVENTING ACCESS OF THE BASAL TRANSCRIPTION.
CC   -!- SUBUNIT: FORMS A LARGE COREPRESSOR COMPLEX THAT CONTAINS SIN3A/B
CC       AND HISTONE DEACETYLASES HDAC1 AND HDAC2. THIS COMPLEX ASSOCIATES
CC       WITH THE THYROID (TR) AND THE RETINOID ACID RECEPTORS (RAR) IN THE
CC       ABSENCE OF LIGAND, AND MAY STABILIZE THEIR INTERACTION WITH TFIIB.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; SMRT/TRAC-2 (SHOWN HERE) AND
CC       TRAC-1; ARE PRODUCED BY ALTERNATIVE SPLICING. TRAC-1 CONTAINS ONLY
CC       THE C-TERMINAL RECEPTOR-INTERACTING DOMAIN AND ACTS AS AN
CC       ANTIREPRESSOR.
CC   -!- TISSUE SPECIFICITY: UBIQUITOUS. HIGH LEVELS OF EXPRESSION ARE
CC       DETECTED IN LUNG, SPLEEN AND BRAIN.
CC   -!- INDUCTION: REGULATED DURING CELL CYCLE PROGRESSION.
CC   -!- DOMAIN: THE N-TERMINAL REGION CONTAINS REPRESSION FUNCTIONS THAT
CC       ARE DIVIDED INTO THREE INDEPENDANT REPRESSION DOMAINS (RD1, RD2
CC       AND RD3). THE C-TERMINAL REGION CONTAINS THE NUCLEAR RECEPTOR-
CC       INTERACTING DOMAINS THAT ARE DIVIDED IN TWO SEPARATE INTERACTION
CC       DOMAINS (ID1 AND ID2).
CC   -!- DOMAIN: THE TWO INTERACTION DOMAINS (ID) CONTAIN A CONSERVED
CC       SEQUENCE REFERRED TO AS THE CORNR BOX. THIS MOTIF IS REQUIRED AND
CC       SUFFICIENT TO PERMIT BINDING TO UNLIGANDED TR AND RARS. SEQUENCES
CC       FLANKING THE CORNR BOX DETERMINE NUCLEAR HORMONE RECEPTOR
CC       SPECIFICITY.
CC   -!- SIMILARITY: CONTAINS 1 SANT-A DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 MYB-LIKE DOMAIN.
CC   -!- SIMILARITY: CONTAINS 2 CORNR BOX.
CC   -!- SIMILARITY: BELONGS TO THE N-COR NUCLEAR RECEPTOR COREPRESSORS
CC       FAMILY.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; AF113003; AAD20946.1; -.
DR   EMBL; AF125672; AAD22973.1; -.
DR   EMBL; U37146; AAC50236.1; -.
DR   EMBL; S83390; AAB50847.1; -.
DR   EMBL; U80750; AAB91446.1; -.
DR   MIM; 600848; -.
DR   InterPro; IPR001005; Myb_DNA_bind.
DR   Pfam; PF00249; myb_DNA-binding; 2.
DR   SMART; SM00395; SANT; 2.
DR   PROSITE; PS50090; MYB_3; 1.
KW   Nuclear protein; Transcription regulation; DNA-binding; Repressor;
KW   Coiled coil; Alternative splicing.
FT   DOMAIN      174    215       COILED COIL (POTENTIAL).
FT   DOMAIN      254    312       INTERACTION WITH SIN3A/B (BY SIMILARITY).
FT   DNA_BIND    429    474       SANT-A (POTENTIAL).
FT   DNA_BIND    613    657       MYB.
FT   DOMAIN      522    561       COILED COIL (POTENTIAL).
FT   DOMAIN      778    820       PRO-RICH.
FT   DOMAIN     2139   2143       CORNR BOX OF ID1.
FT   DOMAIN     2342   2346       CORNR BOX OF ID2.
FT   DOMAIN      494    510       POLY-GLN.
FT   DOMAIN      682    685       POLY-LYS.
FT   DOMAIN      994   1002       POLY-PRO.
FT   DOMAIN     1384   1389       POLY-PRO.
FT   DOMAIN     1842   1846       POLY-GLY.
FT   DOMAIN     2479   2482       POLY-PRO.
FT   VARSPLIC      1   1702       MISSING (IN ISOFORM TRAC-1).
FT   VARSPLIC   2353   2398       MISSING (IN ISOFORM TRAC-1).
FT   CONFLICT      7      7       L -> P (IN REF. 2).
FT   CONFLICT    295    295       K -> E (IN REF. 2).
FT   CONFLICT    309    309       L -> W (IN REF. 2).
FT   CONFLICT    352    352       MISSING (IN REF. 2).
FT   CONFLICT    365    365       A -> P (IN REF. 2).
FT   CONFLICT    612    613       SS -> EF (IN REF. 5).
```

```
FT   CONFLICT    711    711       S -> T (IN REF. 2).
FT   CONFLICT    724    740       MISSING (IN REF. 2).
FT   CONFLICT    787    796       RRTSRAPIEP -> PEDIPAPTES (IN REF. 2).
FT   CONFLICT    804    804       G -> L (IN REF. 2).
FT   CONFLICT    814    814       S -> F (IN REF. 2).
FT   CONFLICT    817    817       A -> S (IN REF. 2).
FT   CONFLICT    889    889       G -> R (IN REF. 2).
FT   CONFLICT   1023   1030       SRSPAPPA -> MEAWDAHP (IN REF. 3).
FT   CONFLICT   1034   1034       A -> AEKPVFFPA (IN REF. 2).
FT   CONFLICT   1894   1894       K -> T (IN REF. 4).
FT   CONFLICT   2494   2494       P -> A (IN REF. 4).
SQ   SEQUENCE   2517 AA;  274031 MW;  F5805C01761258C0 CRC64;


alignment_scores:
            Quality:  114.50              Length:    593
              Ratio:  0.424                Gaps:      33
 Percent Similarity:  45.531   Percent Identity: 21.585

alignment_block:
 US-09-303-518D-465 x NCR2_HUMAN   ..

 Align seg 1/1  to: NCR2_HUMAN  from: 1  to: 2517

      65 CACACGCCTCAGATTTGGCAAACGATTCTTTTATCCGGCAGGTTCTCGAC 114
         |||:::|||   ||||||                        |||   :::||
    1823 HisAlaPro...IleTrp.....................ArgProGlyTh 1831

     115 CGTCAGCATTTCGAACCCGACGGGAAATACCACCTATTCGGCAGCAGGGG 164
         |   :::   |||:::   :::|||:::::      :::::::::|||||
    1831 rGluGlnSerSerGlySerSerGlySerSer......GlyGlyGlyGlyG 1846

     165 GGAACTTGCCGAGCGCAGCGGTCATATCGGATTGGGAAACATACAAAGCC 214
         ||:::            |||      |||      :::   ::::::
    1846 lySerSerSerArgProAlaSerHisSerHisAlaHisGlnHisSerPro 1862

     215 ATCAGTTGGGCAACCTGTTCATCC...AGCAGGCGGCCA.......... 250
         ||||||      |||   ::::::   ::::::||||||
    1863 IleSerProArgThrGlnAspAlaLeuGlnGlnArgProSerValLeuHi 1879

     251 ..........TTAAAGGAAATA.......................... 262
                  :::|||   |||
    1879 sAsnThrGlyMetLysGlyIleIleThrAlaValGluProSerLysProT 1896

     263 ........TCGGCTACATTGTCCGCTTTTCCGATCACGGGCACGAAGTCC 304
                 :::|||   ::::::   |||:::      :::   :::
    1896 hrValLeuArgSerThrSerThrSerSerProValArgProAlaAlaThr 1912

     305 ATTCCCCCTTCGACAACCATGCCT.....................CACAT 333
         :::|||||||:::|||      |||
    1913 PheProProAlaThrHisCysProLeuGlyGlyThrLeuAspGlyValTy 1929

     334 TCCGATTCTGATGAAGCCGGTAGTCCCGTTGACGGATTCAGCCTTTACCG 383
          |||   ||||||:::||||||:::          ::::::      :::|
    1929 rProThrLeuMetGluProValLeuLeuProLysGluAlaProArgValA 1946

     384 CATCCATTGGGACGGATACGAACACCATCCCGCCGACGGCTATGACGGGC 433
         ||                  :::          |||   |||   ||||||
    1946 la.................ArgProGluArgProArgAlaAspThrGly 1956

     434 CACAGGGCGGCGGCTATCCCGCTCCCAAAGGCGCGAGGGATATATACAGC 483
         |||         |||   |||   |||         |||:::      ||
    1957 HisAlaPheLeuAlaLysPro...ProAlaArgSerGlyLeuGluProAl 1972

     484 TACGACATAAAAGGCGTTGCCCAAAATATCCGCCTCAACCTGACCGACAA 533
         |:::::::            |||:::   :::::::::      |||
    1972 aSerSer............ProSerLysGlySerGluProArgPro...L 1984

     534 CCGCAGCACCGGACAACGGCTTGTCGACCGTTTCCACAATACCGGTAGTA 583
           :::   |||   :::|||   :::|||::::::      |||
    1984 euValProProValSerGlyHisAlaThrIleAlaArgThrPro...... 1998

     584 TGCTGACGCAAGGAGTAGGCGACGGATTCAAACGCGCCACCCGATACAGC 633

                              :::   :::   ||||||      :::::
    1999 .....................AlaLysAsnLeuAlaProHisHisAlaSe 2008

     634 CCCGAGCTGGACAGATCGGGCAATGCCGCCGAAGCTTTCAACG....... 676
         :|||         |||         ||||||      ::::::
    2008 rPro.........AspProProAlaProProAlaSerAlaSerAspProH 2022

     677 ..GCACTGCAGATATCGTCAAAAACATCATCGGCGCGGCAGGAGAAATTG 724
                ::   :::||||||      |||   ::::::   :::|||
    2022 isArgGluLysThrGlnSerLysProPheSerIleGlnGluLeuGluLeu 2038

     725 TCGGC.........................GCAGGCGATGCCGTGCAGGG 749
                                       :::   :::::::|||:::
    2039 ArgSerLeuGlyTyrHisGlySerSerTyrSerProGluGlyValGluPr 2055

     750 TATAAGCGAAGGCTCAAACATTGCTGTTATGCACGGCTTGGGTCTGCTTT 799
          :::|||      |||:::   ::::::   |||      ||||||   :
    2055 oValSerProValSerSerProSerLeuThrHisAspLysGlyLeuProL 2072

     800 CCACCGAAAACAAGATGGCGCGCATCAACGATTTGGCAGATATGGCGCAA 849
         ::      :::::::::   ::: |||   :::                ::
    2072 ysHisLeuGluGluLeuAspLys.SerHisLeu..............Gl 2083

     850 CTCAAAGACTATGCCGCAGCAGCCATCCGCGATT......GGGCAGTCCA 893
         :::::::      |||:::||||||:::   :::      |||   :::
    2083 uGlyGluLeuArgProLysGlnProGlyProValLysLeuGlyGlyGluA 2100

     894 AAACCCCAATGCCGCAC............AAGGCATAGAAGCCGTCAGCA 931
           :::   :::||||||                        :::|||||||:::
    2100 laAlaHisLeuProHisLeuArgProLeuProGluSerGlnProSerSer 2116

     932 ATATCTTTACGGCAGTCATCCCCGTCAAAGGGATTGGAGCTGTTCGGGGA 981
             |||   |||:::::::|||:::              :::   |||::
    2117 SerProLeuLeuGlnThrAlaProGly............ValLysGlyHi 2129

     982 AAATACGGCTTGGGCGGCATCACGGCACATCCTGTCA...AGCGGTCGCA 1028
         ::::   :::   :::   ::::::||||||   :::          :::
    2129 sGlnArgValValThrLeuAlaGlnHisIleSerGluValIleThrGlnA 2146

    1029 GATGGGCGAGATCGCATTGCCGAAAGGGAAATCCGCCG............ 1066
           ::::::|||   |||   ::::::   :::   |||
    2146 spTyrThrArgHisHisProGlnGlnLeuSerAlaProLeuProAlaPro 2162

    1067 .....................TCAGCGACAATTTTGCCGATGCGGCATAC 1095
                              |||:::   :::|||   :::|||:::
    2163 LeuTyrSerPheProGlyAlaSerCysProValLeuAspLeuArgArgPr 2179

    1096 GCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGA 1145
          |||:::    :::   ||||||   |||   :::            |||
    2179 oProSer.AspLeuTyrLeuProProProAspHis............Gly 2191

    1146 GCAGCGTTACGGCAAAGAAAACATCACCTCCTCAACCGTGCCGCCGTCAA 1195
         |||       |||         |||
    2192 AlaProAlaArgGlySerProHis........................... 2199

    1196 ACGGAAAGAATGTGAAACTGGCAAACAAACGCCACCCGAAGACCAAAGTG 1245
                   |||   ||||||::::::|||      ::::::::::
    2200 ..........SerGluGlyGlyLysArgSerProGluProAsnLysThrS 2213

    1246 CCGTTTGACGGTAAAGGGTTTCCGAATTTTGAAAAAGACGTAAAATACGA 1295
         ::|||     |||   |||          |||      :::
    2213 erVal.LeuGlyGlyGly............GluAspGlyIle........ 2222

    1296 TACGAGAATTAATACCGCTGTACCACAAGTGAATCCTATAGATGAACCCG 1345
                                                 |||||||
    2223 ........................................GluProV 2225

    1346 TCTTTAATCCTAAAGGTTCTGTCGGATCGGCTCATTCTTGGTCTATAACT 1395
         ||       |||:::|||   :::      :::||||||   |||   :::
    2225 alSerProProGluGlyMetThrGluProGlyHisSerArgSerAlaVal 2241

    1396 GCCAGAATTCAATACGCAAAATTACCAAGGCAAGGTAGAATCAGATATAT 1445
                  |||      |||   |||:::|||
```

```
2242 ............TyrProLeuLeuTyrArgAspGly.............. 2249

1446 CCCACCTAAAAATTACTCTCCTTCAGCACCGCTACCAAAAGGACCTAATA 1495
           ::::::   :::||||||           |||:::|||:::|
2250 .......GluGlnThrGluProSerArgMetGlySerLysSerProGlyA 2264

1496 AT..............GGATATTTGGATAAATTTGGTAATGAATGGACT 1530
     ||              ::::::::::::|||:::   :::|||   :::
2264 snThrSerGlnProProAlaPhePheSerLysLeu...ThrGluSerAsn 2279

1531 AAAGGTCCATCAAGAACTAAAGGTCAAGAATTTGAATGGGATGTTCAATT 1580
     ::::::      ::::::|||   ||||||:::
2280 SerAlaMetValLysSerLysLysGlnGluIle................. 2290

1581 GTCTAAAACAGGAAGAGAGCAACTTGGATGGGCTAGTAGGGATGGTAAGG 1630
                :::::::::|||:::      :::|||:::      :
2291 .............AsnLysLysLeuAsnThrHisAsnArgAsnGluProG 2303

1631 ATTTAAATATATCAATTGATGGA 1653
     ::   |||||||||      |||
2303 luTyrAsnIleSerGlnProGly 2310
```